informatics

MDPI

Article

A Knowledge-Distilled Conformer Framework for On-Device Sleep Assessment

Geetha Lekshmy Vanajamba [1] and Simi Surendran [2,*]

[1] Department of Computer Science and Applications, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India; geethalekshmy@am.amrita.edu
[2] Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Amritapuri, India
* Correspondence: simisurendran@am.amrita.edu

Abstract

Sleep disorders are prevalent across the population and, in older adults, are associated with cognitive decline, reduced functional capacity, and diminished quality of life. While polysomnography (PSG) provides detailed neurophysiological information for sleep assessment, its reliance on specialized expertise and high computational cost prevents the scalable, continuous monitoring needed in home and long-term care settings. We propose an edge-deployable, Conformer-based multimodal PSG framework for automated sleep staging. A high-capacity teacher model encodes electroencephalogram (EEG), electrooculogram (EOG) and electromyogram (EMG) through per-channel Conformers, a graph transformer and cross-attention fusion, with a temporal transformer capturing inter-epoch dynamics. To enable low-power operation, a dual knowledge-distillation strategy combining feature and logit-level transfer compresses the teacher into a compact, EEG-only student. The framework is evaluated across two clinical cohorts, the Multi-Ethnic Study of Atherosclerosis (MESA) and the Cleveland Family Study (CFS). The teacher achieves performance competitive with state-of-the-art methods on both cohorts, but distillation retention is cohort-dependent. On CFS, the student retains 92.7% of the teacher's macro-F1, with above 90% retention for Wake, N1, N2, and N3, though REM retention is limited to 79.7%. The distilled student reduces parameters by 38.5× and latency by 11.3×, processing a 30-s epoch in 6.64 ms using 0.64 MB on an edge-class CPU. On the older, sparser MESA cohort, retention falls to 78.1% of the teacher's macro-F1, with REM retention dropping to 52.1%. These results show that resource-efficient multimodal sleep staging is feasible for on-device deployment, while identifying robust minority-stage retention on sparser, older cohorts as the key remaining challenge for reliable cross-cohort operation.

Keywords: sleep staging; knowledge distillation; conformer; edge AI; polysomnography; multimodal fusion; graph transformer; cross attention fusion; on-device inference; model compression

Check for updates

Academic Editor: Jiang Bian

Received: 22 June 2026
Revised: 20 September 2026
Accepted: 26 September 2026
Published: 8 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Sleep plays a critical role in maintaining cognitive function and emotional well-being. The strong connection between sleep disruption, disease, and mortality has made the diagnosis and mitigation of sleep disorders a clinical priority [1]. Sleep disturbances increase with aging and causes poor sleep quality, risk of sleep apnea, and insomnia symptoms [2]. Such disturbances often precede the clinical diagnosis of dementia by

several years, and the rhythmicity of the human circadian system weakens with age [3]. Insomnia symptoms associated with cognitive frailty are particularly prevalent among elderly women [4]. Sleep-health assessment is therefore a valuable tool for the early detection of accelerated cognitive decline and for timely intervention [5]. This motivates the need for continuous, scalable sleep monitoring in the populations most at risk, including in home and long-term care settings [6].

Polysomnography (PSG) remains the standard for diagnosing sleep disorders and records multimodal physiological signals, including electroencephalogram (EEG), electrooculogram (EOG), and electromyogram (EMG). Following the American Academy of Sleep Medicine (AASM) guidelines [7], these signals are scored in 30-s epochs into five stages: Wake, N1, N2, N3, and rapid eye movement (REM). Identification of these stages is essential for detecting pathological patterns such as sleep fragmentation, which is strongly associated with cardiovascular disease, metabolic dysfunction, and neurocognitive decline [8]. However, manual sleep scoring is effort-intensive, slow, and prone to inter-scorer inconsistency, motivating the development of automated, reliable, and scalable sleep-staging systems.

Recent advances in deep learning have improved automated sleep staging. Convolutional neural networks have been widely adopted to learn temporal and spectral EEG representations [9–14], recurrent and transformer-based architectures further model long-range inter-epoch dependencies [15–18]. Convolution-augmented transformers (Conformers) have proven particularly effective in jointly capturing local and global structure in EEG and related biosignals [19–21]. Although many methods operate on a single EEG channel, multimodal approaches that combine EEG with EOG and EMG have been shown to improve staging [22–25]. In parallel, a distinct line of work targets deployability. Lightweight and knowledge-distilled models reduce the parameter and computational cost of on-device inference [26–29], while alternative-sensor methods based on photoplethysmography or cardiorespiratory signals enable wearable home monitoring [30,31].

Despite this progress, a fundamental trade-off persists. The most accurate sleep-staging models exploit rich multimodal PSG but are too large and computationally heavy for continuous on-device use, whereas deployable models achieve efficiency by restricting themselves to a single EEG channel or to indirect photoplethysmographic and cardiorespiratory signals, sacrificing multimodal neurophysiological information. Indirect-sensor methods are also known to degrade with increasing age and apnea severity, conditions most prevalent in older populations, and often require continuous full-night recordings.

Among compression techniques, pruning and quantization can reduce model size but often degrade clinical accuracy, whereas knowledge distillation (KD) retains more of the teacher's representational quality. KD, introduced by Hinton et al. [32], transfers knowledge from a teacher model to a student model by training the student on both the teacher's soft probabilistic outputs and the hard ground-truth labels. However, transferring knowledge between a heterogeneous multimodal teacher and a compact student remains underexplored [26,33]. Building on this, we identify three persisting challenges: (i) CNN-based architectures may inadequately capture global temporal dependencies and inter-channel interactions; (ii) multimodal fusion is often performed through simple concatenation rather than adaptive attention-based mechanisms, limiting cross-modal contextual modeling; and (iii) knowledge distillation faces the teacher-student capacity gap and the difficulty of transferring knowledge between structurally dissimilar models [26,33].

To address these gaps, we propose an edge-deployable, multimodal Conformer-based framework that retains the diagnostic richness of full PSG while operating within the resource constraints of home-based and long-term care settings.

The main contributions of this work are summarized as follows:

- A unified multimodal teacher-student framework for deployable sleep staging, combining Conformer-based temporal learning, graph-based inter-channel modeling, cross-attention fusion, and dual knowledge distillation.
- A Conformer-graph encoder that captures intra-epoch EEG features and models inter-channel neurophysiological dependencies using Conformer-derived node representations.
- A tri-modal cross-attention fusion that integrates EEG, EOG, and EMG representations through attention-based cross-modal correspondence rather than concatenation or late ensembling.
- A dual knowledge-distillation strategy that compresses the multimodal teacher into a compact, EEG-only student suitable for low-power edge deployment.
- A cross-modal hallucination mechanism that enables EEG-only inference by synthesizing proxy EOG and EMG representations from EEG, guided by the multimodal teacher's embeddings through feature-level distillation.

The remainder of this paper is organized as follows. Section 2 reviews related work on knowledge distillation, lightweight architectures, multimodal fusion, and graph-based modeling for sleep staging. Section 3 describes the proposed teacher and student models. Section 4 presents the experimental results, comprising the ablation study, per-cohort staging performance, teacher-student retention and teacher-ceiling analyses, and comparison with state-of-the-art methods. Section 5 concludes the paper and outlines directions for future work.

2. Related Works

This section summarizes the state of the art relevant to the proposed framework across four areas: knowledge distillation for model compression, lightweight architectures for efficient sleep staging, multimodal fusion of physiological signals and graph-based modeling of inter-channel dependencies.

Gou et al. [34] introduced an attention-map based correlation for distillation, outperforming prior techniques with 78.87% accuracy on CIFAR-100 using ResNet teacher–student pairs. Chen et al. [35] proposed a knowledge-review mechanism in which multiple teacher layers supervise a single student layer, improving classification and object detection. A prediction-augmentation strategy [36] converted single teacher and student output into multiple temperature-scaled predictions aligned at the instance, class, and batch levels, achieving the highest accuracy among logit-distillation methods for homogeneous teacher–student architectures on CIFAR-100. These foundational methods establish KD as an effective route to compact models, motivating its adoption for resource-constrained sleep staging.

Within sleep staging, Kim et al. [37] proposed cross-model distillation between a self-attention CNN teacher and a transformer student, improving stage classification by 1.7% over state-of-the-art models on Sleep-EDF. However, the study uses a single channel and a single dataset, leaving multi-channel scalability and cross-dataset generalization unaddressed. Zhang et al. [38] distilled across domains and channels: a bidirectional RNN pre-trained on multiple MASS channels, initialized both teacher and student, which were then trained on Sleep-EDF with a multi-channel teacher and a single-channel student, achieving 86.5% accuracy in single-channel staging comparable to multi-channel performance. Most closely related to the present work, Ziyu et al. [26] addressed distillation between structurally dissimilar architectures by projecting the student's features into a logit-sized space and measuring output divergence via KL divergence, with a teacher-assistant module moderating the capacity gap. Their CNN–transformer teacher and lightweight

student achieved 85.4% and 81.6% accuracy (teacher) and 84.7% and 80.0% (student) on Sleep-EDF and ISRUC, respectively. This work was unimodal, whereas the present study distills across a structurally heterogeneous multimodal teacher. Finally, Wang et al. [39] distilled a SleePyCo-based teacher into a transformer-encoder student using combined distillation and prediction losses, reaching 82.4% accuracy on Sleep-EDF.

Another research direction is in the compact architecture design. Tan et al. [40] proposed a mobile neural architecture search that jointly optimizes accuracy and inference latency over a hierarchical search space. The resulting baseline achieved 75.2% accuracy with 3.9 M parameters and 78 ms latency on ImageNet, improving on the manually designed MobileNet. For sleep staging specifically, a single-channel lightweight model [41] combined raw time-domain signals with their amplitude spectrum as a frequency-domain input, using convolutional and self-attention layers to reach 84.8% accuracy, a Cohen's k of 0.787, and 0.47×10^6 parameters on Sleep-EDF-78. A multi-channel network [42] took time–frequency images of EEG channels as input, applying depthwise-separable convolutions for feature extraction and a convolutional block attention module for inter-channel modeling, achieving 88.2% accuracy, k of 0.84, and macro-F1 of 82.4% on Sleep-EDF-20. An attention-based lightweight model [43] used a multi-resolution CNN combining pointwise and depthwise convolutions for feature extraction and multi-head attention for temporal modeling, reporting 86.6% accuracy and 79.2% macro-F1 on Sleep-EDF with 55.3 MFLOPs and 180 K parameters. These models demonstrate that efficient inference is achievable, but they operate predominantly on single-channel EEG, not considering the complementary information available from multimodal recordings.

Although many automated methods rely on single-channel EEG, combining EEG with EOG and EMG improves staging, as these modalities capture complementary information essential for stages such as REM. Yubo et al. [22] proposed MMASleepNet, which integrates multiple modalities through a multi-head attention module rather than simple concatenation. Phan et al. [23] jointly modeled EEG and EOG for sequence-to-sequence staging. More recent work has employed cross-attention mechanisms for feature-level inter-modal interaction [24,44], motivated by the finding that naive concatenation insufficiently models cross-channel correlations. Wang et al. [45] proposed MASleepNet, combining multi-scale convolution and BiLSTM with attention mechanisms to fuse EEG, EOG and EMG, achieving an accuracy of 84.53% Sleep-EDF-20 dataset. Nevertheless, most multimodal approaches are not designed for resource-constrained settings, and the integration of multimodal fusion with model compression for on-device deployment remains largely unaddressed.

To capture spatial relationships among EEG channels, several works represent the electrode montage as a graph. Jia et al. [46] introduced GraphSleepNet, which adaptively learns an adjacency matrix among channels for spatial–temporal sleep staging, and subsequent work has extended this to graph-transformer formulations incorporating structural and positional encodings [47]. These approaches typically construct graph nodes from comparatively shallow per-channel features. In contrast, the present work derives node representations from full per-channel Conformer encoders, supplying richer node features to the graph transformer, and embeds this component within a multimodal, distillable pipeline rather than treating it as a standalone classifier.

Although existing studies have made significant progress in knowledge distillation, lightweight sleep-staging models, and multimodal learning, these approaches have been studied largely independently. Distillation and lightweight studies mostly on single-channel EEG and do not fully exploit other physiological signals. Multimodal and graph-based methods improve accuracy but are computationally intensive and not suitable for edge deployment. Furthermore, the use of knowledge distillation between heterogeneous multimodal teacher and student architectures remains relatively unexplored and evaluation

is typically confined to a single benchmark dataset. The present work addresses these issues by a unified end-to-end framework distilling a heterogeneous multimodal Conformer–Graph teacher into an edge-deployable student and validating it across two clinical cohorts.

3. Materials and Methods

3.1. Datasets

We evaluate the proposed framework on two clinical PSG cohorts accessed through the National Sleep Research Resource (NSRR) [48,49]. Both cohorts comprise predominantly older adults, consistent with the target setting of home-based and long-term care monitoring.

The Multi-Ethnic Study of Atherosclerosis (MESA) sleep dataset [50,51] is a large, multi-center cohort containing PSG and health data from diverse, primarily older adults, used to study cardiovascular risk and sleep disorders. The MESA cohort includes participants aged between 45 and 84 years who were enrolled to investigate prevalence of cardiovascular diseases (CVD) and identify CVD risk factors. After applying our inclusion criteria (recordings with valid EEG, EOG, and EMG channels and complete stage annotations), 2057 PSG recordings (21,741 h) are used in this work. The PSG recording montage captured a range of physiological signals, including multiple EEG channels (central, occipital, and frontal), EOG, EMG, plethysmography, ECG, and pulse oximetry, of which the EEG, EOG, and EMG signals are specifically used in this work. Figure 1 represents two complementary views of sleep stage composition across 2057 PSG recordings (21,741 h) from the MESA dataset. Bar chart on the left panel of the figure illustrates domination of wake stage (43%), followed by N2 (33%), REM (10%), N1 (8%), and least represented stage being N3 (6%). Right panel shows the distribution of each stage's percentage within individual recordings. Wake stage has IQR of roughly 30–50%, showing variable sleep efficiency across subjects. N2 seems to be consistent whereas N1, N3 and REM are highly variable.

The Cleveland Family Study(CFS) [50,52] is a family-based study of sleep apnea, having 2284 individuals from 361 families. PSG recording has 14 channels, including EEG, EOG, EMG, pulse oximetry, oronasal flow, plethysmography. The CFS visit-5 provides 726 recordings out of which we have included only 413 from participants aged 40 and above. There are 138 families in the 413 recordings. Figure 2 shows the sleep-stage composition of the CFS recordings. Out of 487,516 epochs (4062.6 h), Wake (40.4%) and N2 (35.9%) together account for roughly 76% of all epochs. Severe class imbalance is evident: N1 is the least represented stage at 3.3%, with N3 (9.2%) and REM (11.2%) also sparsely represented. The box plot shows that Wake has the highest spread, and its upper outliers indicate that some recordings are dominated by wake. N2 also shows a wide distribution, with a median near 36%. N3 and REM have moderate spread, and their scattered outliers reflect participant-to-participant differences in deep-sleep and REM proportion.

Both datasets provide recordings in European Data Format (.edf) with .xml annotation files containing sleep-stage labels and event markers. Sleep scoring standard is different for both cohorts. MESA follows AASM guidelines (five stages: Wake, N1, N2, N3, REM), whereas CFS uses the older Rechtschaffen and Kales (R&K) standard, which additionally distinguishes sleep stages S3 (Stage 3) and S4 (Stage 4). To maintain a uniform label space across cohorts, R&K stages S3 and S4 are merged into a single N3 (slow-wave) stage, yielding the five AASM classes (Wake, N1, N2, N3, REM) for both datasets. Table 1 compares the two cohorts. Three EEG channels, one EOG channel, and one EMG channel are selected for building the automated staging model.

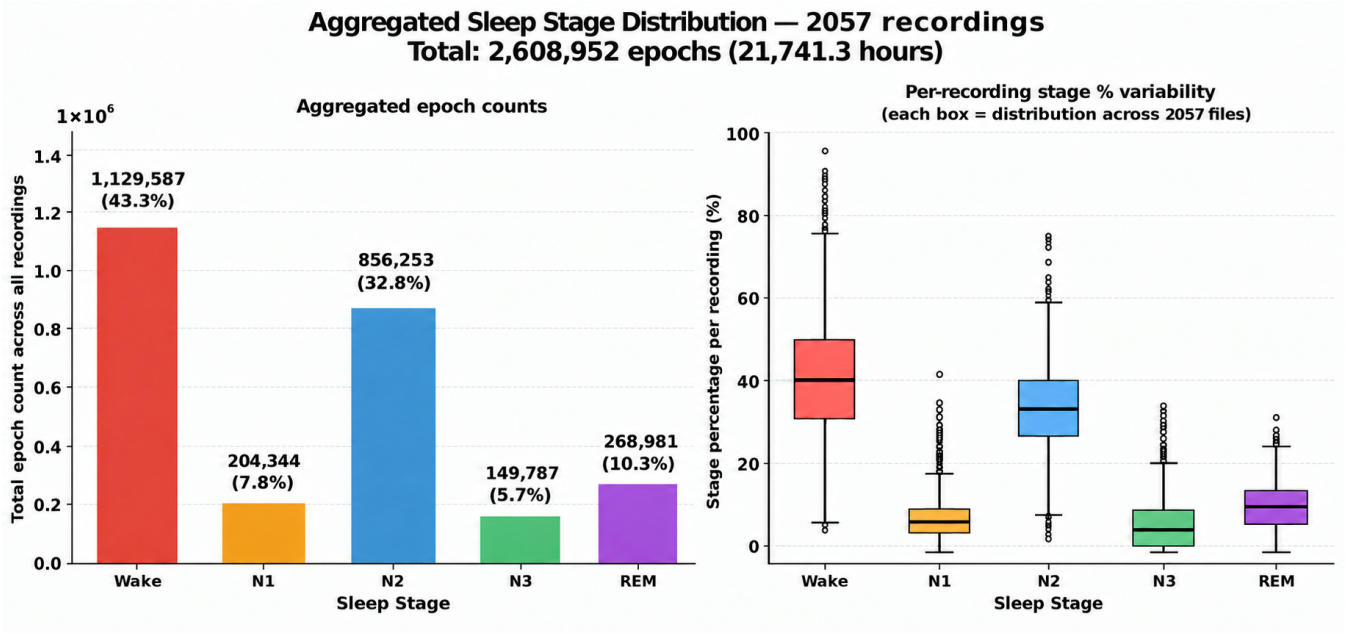


Figure 1. Aggregated Sleep Stages in MESA—Total 30 s epoch counts, per sleep stage variability.

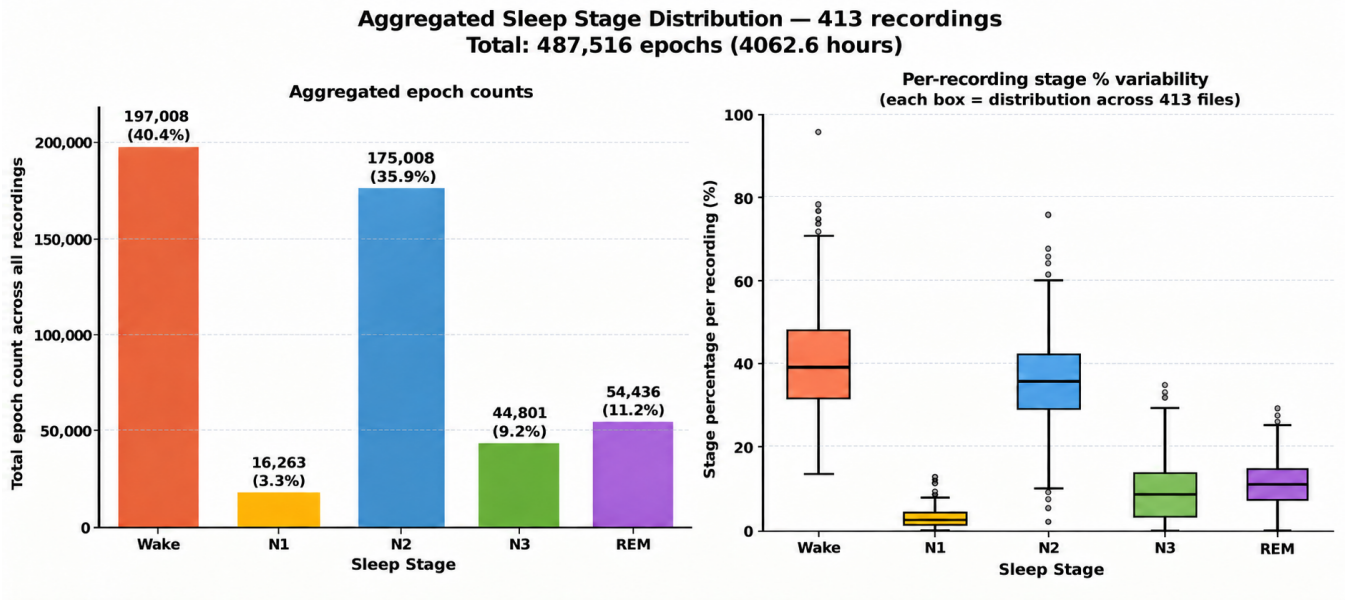


Figure 2. Aggregated Sleep Stages in CFS—Total 30 s epoch counts, per sleep stage variability.

Table 1 shows basic comparison of MESA and CFS with respect to participant number, sampling rate, channels and annotation standard. Three EEG channels, one EOG and one EMG channels are selected for building the automated sleep staging model.

Table 1. Comparison of the MESA and CFS datasets.

Aspect	MESA [50,51]	CFS [50,52]
Number of Recordings	2057	413
Sampling Rate (EEG, EOG, EMG)	256 Hz	128 Hz
Channels [a] : EEG	Fz–Cz, Cz–Oz, C4–M1	C3–Fz, C4–Fz, A1–Fz
: EMG	chin EMG	EMG1–Fz (Center Chin EMG)
: EOG	EOG-L – Fpz (Left EOG)	LOC – Fpz (left EOG)
Annotation Standard	AASM	R&K

[a] Channels used in this work are listed out for both MESA and CFS datasets.

3.2. Multimodal Teacher–Student KD Framework

The proposed framework, illustrated in Figure 3, comprises three stages: data preprocessing, a high-capacity multimodal teacher, and a compact distilled student. The teacher encodes EEG, EOG, and EMG through per-channel Conformers, models inter-channel dependencies with a graph transformer, fuses the modalities through cross-attention, and captures inter-epoch context with a temporal transformer. The student is a lightweight, EEG-only model that internally reconstructs the EOG and EMG representations and is trained to match the teacher through combined feature and logit-level knowledge distillation. The student requires only EEG signals during inferencing, EOG and EMG are needed only during training. Each stage is described below.

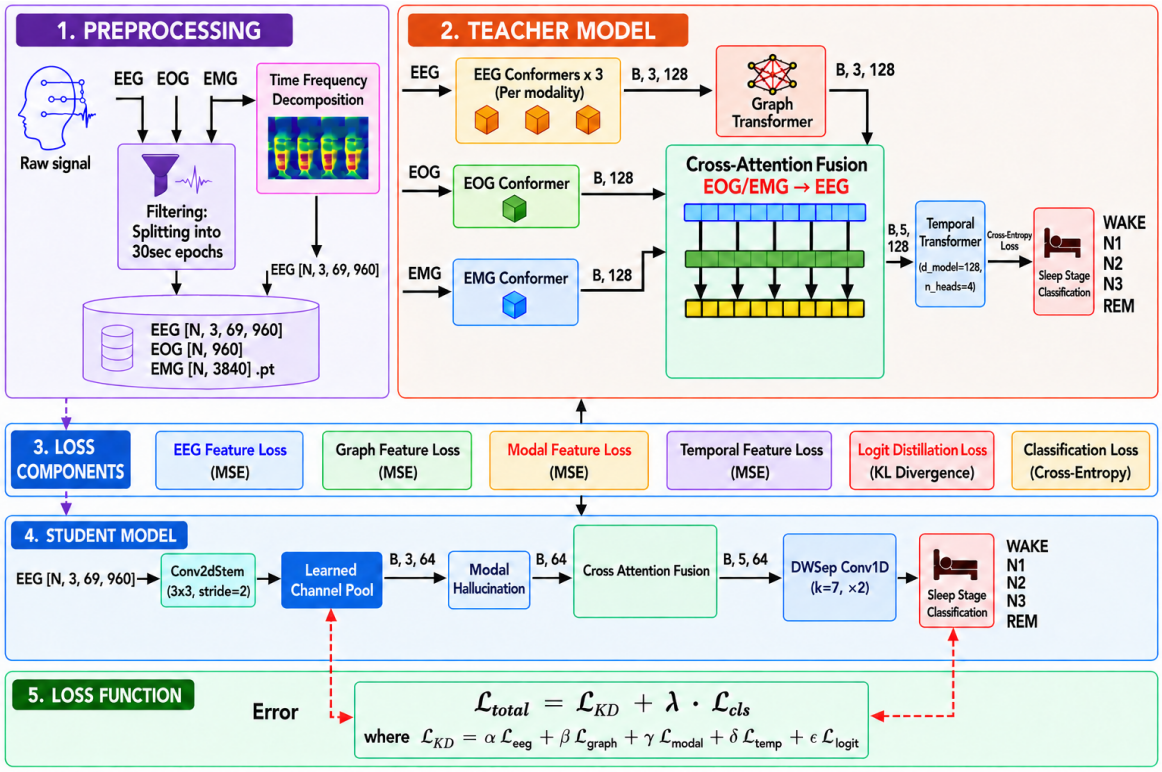


Figure 3. Proposed Teacher –Student framework for sleep stage classification. A Conformer–Graph Transformer teacher guides a lightweight student through hallucinated modality supervision and multi-level knowledge distillation.

3.2.1. Data Pre-Processing

Each raw .edf recording is loaded, and initial wake stages are removed to reduce the class imbalance due to excessive wake stages in samples. The last 60 s of the initial wake stages are preserved to prevent the loss of stage transition information from wake stage to light sleep. The corresponding Wake annotations in the xml file are trimmed identically to avoid misalignment of the labels. Five channels consisting of 3 EEG channels, 1 EOG and 1 EMG channels are selected from the PSG data. Band-pass filter of 0.3–40 Hz, 0.3–35 Hz and 10–60 Hz are applied to EEG, EOG and EMG signals respectively. All signals are resampled to 128 Hz, and normalised using Z-score normalisation. Annotation file is parsed to extract sleep stages and its duration. The resampled signals are divided into 30 s epochs and assigned a sleep stage label extracted from annotation file. CFS uses R&K scoring, so to maintain uniformity across two datasets four sleep stages (N1, N2, N3, REM) and Wake are considered in this work.

EEG signals after segmenting into 30 s epochs, are transformed using Continuous Wavelet Transform (CWT) to obtain multi-scale time–frequency representations. CWT captures fine, high-frequency features at small scales and broad, low-frequency trends at large scales [53]. It has adaptive time frequency resolution that helps to capture slow rhythmic patterns and rapid transients, which is ideal for non stationary signals like EEG. The CWT is computed using the Morlet wavelet with nv(voices per octave) = 8, producing a scalogram of 69 frequency scales spanning 0.5–30 Hz, which covers all sleep-relevant EEG frequency bands from delta to beta. Three EEG signals in time-frequency domain and EOG signals are downsampled to reduce storage requirement. EMG signals are not downsampled to preserve high frequency muscle signals. 30 s segmented five channel signals together with the corresponding sleep stages are stored as tensors for further processing.

3.2.2. Conformer–Graph Transformer Teacher

Sleep EEG contains two complementary pattern types. Some are localized waveforms at fixed time scales such as sleep spindles and K-complexes that occupy a narrow temporal window. Others, such as the low-amplitude mixed-frequency activity are locally ambiguous and distinguishable only from surrounding context. Convolutions capture local structure, while self-attention captures long-range context, hence the Conformer [54] combines both in a single block and has outperformed CNN and transformer baselines in speech [54] and vision [55]. The teacher therefore adopts Conformer encoders for all modalities. For per-channel EEG encoding, each of the three preprocessed EEG channels that is stored as a $[B, 3, 69, 960]$ tensor is flattened and divided into patches of size 320, each linearly projected to the model dimension $d = 128$, yielding a sequence of length 207. A sinusoidal positional encoding is added. There are three Conformer blocks, each one has two half-step feed forward modules that sandwich attention and convolution modules. Multi-head self attention module has 4 heads, preceded by normalisation layer. A depthwise Conv1D with kernel size 31 is included in the convolution module. The output from three Conformer blocks corresponding to three EEG signals are next passed into Graph Transformer to model inter channel relationship.

The inter-channel modeling is done through a dynamic graph transformer that treats each channel as a node carrying a 128-dimensional feature. An edge MLP scores every channel pair by concatenating their embeddings and predicting a connection strength, producing a 3×3 adjacency matrix. This adjacency matrix is injected directly into multi-head self-attention as a bias, so strongly connected channel pairs attend to each other more than weakly connected ones. A feedforward layer then refines each node embedding, and finally all three node representations are mean-pooled into a single 128-dimensional vector that captures cross-channel EEG relationships for that epoch.

Multiple modalities EEG, EOG, and EMG, encoded by their own Conformers are combined by a cross-attention fusion module in which the EEG graph embedding forms the query and the EOG and EMG embeddings form the keys and values, producing a fused multimodal embedding for each epoch. The graph-encoded EEG embedding is treated as a single query token, $Q \in \mathbb{R}^{B\times 1\times d}$, while the EOG and EMG embeddings are stacked into a two-token sequence, $K, V \in \mathbb{R}^{B\times 2\times d}$, used as keys and values. Multi-head cross-attention with four heads computes softmax $(QK^{\top}/\sqrt{d_h})V$, where $d_h = d/4$ for each head, the attention distributes weight across the EOG and EMG tokens according to their relevance to the current EEG representation. The attended output is added to the EEG embedding through a residual connection followed by layer normalization, yielding the fused multimodal embedding for the epoch. This design lets the model learn how much eye-movement (EOG) and muscle (EMG) information each epoch's EEG representation

should incorporate, reflecting the clinical reliance on EOG for REM detection and EMG for muscle-atonia assessment. A 30 s epoch has information from all three modalities and is computed for 20 (sequence length) epochs which are then stacked and passed to temporal transformer.

Temporal transformer models the relationship between consecutive epochs, since the stage of one epoch depends on its surrounding context. For instance, sleep stages N1 and REM have low amplitude mixed frequency EEG signals, but they differ in context. So to classify a 30-s epoch as N1 or REM, adjacent epochs should be looked into. In the proposed architecture the 128 dimensional output from Temporal Transformer is given to the classification head, which receives all 20 epoch embeddings $[B, S, 128]$ and outputs one sleep stage label per epoch $[B, S, 5]$ Wake, N1, N2, N3, or REM for each 30-s window in a sequence of length 20.

3.2.3. Student Architecture and Knowledge Distillation

The teacher is trained on three modalities (EEG, EOG, EMG), whereas the student operates on EEG alone at inference and is trained to mimic the teacher's representations and output distributions. We adopt a heterogeneous-model distillation strategy following [26], addressing the teacher-student dimensionality gap with a projector that maps student features into the teacher's feature dimension. The student model is shown in Figure 3. For EEG Encoding, each EEG channel is processed by a Conv2D. This replaces the Conformer encoder of the teacher model. Each stem applies two convolutional layers ($1 \to 16 \to 32$ channels, kernel size 3, stride 2, each followed by batch normalization and a GELU activation), a fixed window average pooling that reduces the feature map to a 4×4 spatial grid, and a linear projection with layer normalization to the student dimension $d = 64$.

The channel aggregation is performed through a learned channel-pooling layer, replacing the teacher's graph transformer. It consists of three learnable scalar weights, used to find weighted sum of three channel embeddings to a single vector. The output of this layer is a 64-dimensional vector, which is a weighted average of three EEG channels.

A cross-modal hallucination layer synthesizes proxy EOG and EMG representations from the EEG signal. This layer has two independent single-layer MLPs which produces two outputs from the 64 dimensional EEG vector. These two outputs are $e\hat{o}g$ and $e\hat{m}g$ (generated EOG, EMG) and are forced to match the real EOG and EMG embedding produced by teacher using a loss function. During training, the real EOG and EMG signals are passed through the teacher's EOG and EMG Conformers, and the resulting embeddings supervise the hallucinated ones through a feature-matching loss, forcing the student to reconstruct genuine multimodal features from EEG alone.

The resultant embeddings from modal hallucination layer are passed into cross attention fusion layer in student which has half the teacher's width and half its heads ($d_{\text{model}} = 64$, Attention heads = 2). EEG embedding forms the query and synthetic EOG, EMG forms the keys/values, so attention flows in a single direction (EEG attending to the hallucinated modalities). The temporal modeling in student is done by a convolution block consisting of two stacked depthwise separable 1D convolution blocks. Each block applies a depthwise convolution with kernel size 7 followed by pointwise convolution, a residual connection and layer normalisation. A classification head follows that takes temporal sequence and predicts stage for each epoch.

3.3. Loss Function

The student is trained with a total objective combining a supervised task loss $\mathcal{L}_{\text{T}}$ and a knowledge-distillation loss $\mathcal{L}_{\text{KD}}$:

$$\mathcal{L}_{\text{total}} = \lambda \mathcal{L}_{\text{T}} + \mathcal{L}_{\text{KD}}$$

where $\mathcal{L}_{\mathrm{T}}$ is the cross-entropy loss. The distillation loss aggregates four feature-level mean squared error (MSE) terms and a logit-level term:

$$\mathcal{L}_{\mathrm{KD}} = \alpha\,\mathcal{L}^{\mathrm{EEG}} + \beta\,\mathcal{L}^{\mathrm{Graph}} + \gamma\,\mathcal{L}^{\mathrm{Modal}} + \delta\,\mathcal{L}^{\mathrm{Temp}} + \varepsilon\,\mathcal{L}_{\mathrm{logit}}$$

Each feature term is computed after projecting the student feature into the teacher's feature dimension:

- $\mathcal{L}^{\mathrm{EEG}}$ matches the student CNN2D stem output to the teacher's per-channel EEG Conformer output,
- $\mathcal{L}^{\mathrm{Graph}}$ matches the student's channel-pooled embedding to the teacher's graph-transformer output,
- $\mathcal{L}^{\mathrm{Modal}}$ matches the hallucinated $\hat{z}_{\mathrm{EOG}}$ and $\hat{z}_{\mathrm{EMG}}$ to the teacher's EOG and EMG Conformer outputs,
- $\mathcal{L}^{\mathrm{Temp}}$ matches the student and teacher temporal-encoder outputs.
- $\mathcal{L}_{\mathrm{logit}}$, the logit term is the temperature-scaled Kullback–Leibler divergence between the teacher and student stage distributions:

$$\mathcal{L}_{\mathrm{logit}} = T^2\,D_{\mathrm{KL}}\big(\sigma(z^t/T)\,\|\,\sigma(z^s/T)\big),$$

where z^t and z^s are the teacher and student logits, σ is the softmax, and T is the distillation temperature. The loss weights and temperature are listed in Table 2. λ weights the supervised task loss, α, β, γ, and δ weight the feature-level distillation terms, ε weights the logit-level term, and T is the softmax temperature.

Table 2. Distillation loss weights and temperature.

Symbol	Loss Term	Value
λ	Stage cross-entropy (task)	0.7
α	EEG feature MSE	0.5
β	Graph feature MSE	0.3
γ	Modal (hallucination) MSE	0.8
δ	Temporal feature MSE	0.4
ε	Logit KD	1.0
T	Distillation temperature	4.0

3.4. Experimental Setup

All experiments were conducted on a high-performance computing system equipped with an NVIDIA H100 GPU with 93 GB RAM, using PyTorch version 2.7.1. Datasets were partitioned at the subject level into 60% training, 20% validation, and 20% test sets to prevent subject leakage. The trimming of initial Wake stages was done for all three data partitions. Each recording was segmented into sequences of $S = 20$ consecutive 30-s epochs overlapping with stride $S/2$ during training, and non-overlapping for validation and test so that every epoch is evaluated once. Three seeds 1, 7 and 42 were used to govern the splits in three different runs of teacher and student across CFS and MESA data. In CFS cohort, our current partition is at the recording (subject) level and does not explicitly group family members. The partition assigns 249 of 413 recordings to training and 82 each to validation and test. 139 of the 164 evaluation recordings (84.8%) come from families that also appear in training. To assess family-level leakage we additionally evaluated CFS under a family-grouped split in which no family appears in more than one partition. Performance remained comparable to the standard split, indicating that our results are not driven by family-level leakage.

The teacher uses an embedding dimension $d_{\text{model}} = 128$ with 4 attention heads. Each modality is encoded by a Conformer of 3 blocks with a feed-forward expansion factor of 4, the three per-channel EEG Conformers use a depthwise convolution kernel of size 31, while the EOG and EMG Conformers use a kernel of size 15. Inter-channel dependencies are modeled by a graph transformer with a learned edge-weight adjacency EEG-graph. EOG, and EMG embeddings are combined through cross-attention fusion in which the EEG-graph embedding forms the query and the EOG and EMG embeddings form the keys and values. A two-layer temporal transformer feed-forward dimension $2\,d_{\text{model}}$, with learnable positional embeddings models inter-epoch context. Both the teacher and student were optimized with AdamW under a cosine-annealing learning-rate schedule whose annealing period was set to the maximum number of training epochs, with a batch size of 16, a dropout rate of 0.01, and a maximum of 50 epochs with early stopping. The student is additionally optimized against a supervised stage loss, logit-level distillation, and feature-level distillation from the frozen teacher. The training configuration is summarized in Table 3.

Table 3. Training configuration for the teacher and student models.

Parameter	Teacher	Student
Optimizer	AdamW	AdamW
LR schedule	Cosine annealing	Cosine annealing
Learning rate	3×10^{-5}	1×10^{-4}
Weight decay	1×10^{-3}	1×10^{-3}
Batch size	16	16
Max epochs	50	50
Early-stopping patience	5	5
Dropout	0.01	0.01
Random seeds	1, 7, 42	1, 7, 42
Sequence length S	20	20
d_{model}	128	64
Attention heads	4	2 [a]

[a] No. of attention heads in Cross Attention Fusion in Student model.

4. Results and Discussion

We report results of the proposed framework on MESA and CFS cohorts. Two labeling schemes were adopted, first standard five-class AASM scheme with classes Wake, N1, N2, N3, REM and second, the four class scheme in which N1 and N2 are merged into a single Light-sleep class. First scheme serves as primary protocol and is used in ablation study, the per-stage teacher/student analysis and the teacher-ceiling analysis. The second scheme is used for comparison with prior state-of-the-art methods, whose results are recorded in four class form. This scheme ensures a fair comparison aligning the label space. Every table and figure caption indicates the scheme it uses and all results are reported as mean $\pm$ standard deviation over three seeds (1, 7, and 42). Performance is assessed using accuracy, macro-F1, weighted-F1, and Cohen's kappa. In this section, we first present performance evaluation of the proposed architecture followed by a discussion. In performance evaluation section an ablation study isolating the contribution of each architectural component, per-stage and overall staging performance of the teacher and distilled student on CFS and MESA datasets, a teacher–student retention analysis, comparison of the framework against recent state-of-the-art methods and finally model complexity are presented. Discussion section interprets the staging results and their deployment implications.

4.1. Performance Evaluation

Table 4 reports an ablation study on the CFS cohort under the five-class scheme, in which each component is added incrementally. A unimodal configuration using only EEG with the Conformer, graph transformer, and temporal transformer attains 74.01% accuracy, Cohen's κ of 0.6457, and macro-F1 of 0.6835. Extending the input to all three modalities (EEG + EOG + EMG), fused by simple aggregation without cross-attention, improves performance to 76.01% accuracy, κ 0.6670 and macro-F1 decreases slightly 0.6711. Replacing the simple aggregation with cross-attention fusion yields the full model, which achieves 77.51% accuracy, κ 0.6920, and macro-F1 0.7050. The multimodal extension provides the largest single gain in accuracy and k, confirming that ocular (EOG) and muscular (EMG) signals carry discriminative information that EEG alone does not recover, while cross-attention fusion provides the largest macro-F1 improvement.

Table 4. Ablation study on the CFS cohort (five-class) evaluating the contribution of each architectural component.

Model Variant	Accuracy (%)	Cohen's κ	Macro-F1
EEG only (Conformer + Graph + Temporal)	74.01 ± 0.9800	0.6457 ± 0.0175	0.6835 ± 0.0145
EEG + EOG + EMG (+ Graph + Temporal)	76.01 ± 1.1000	0.6670 ± 0.0110	0.6711 ± 0.0131
Full Model	77.51 ± 1.870	0.6920 ± 0.0212	0.7050 ± 0.0098

The four-class staging results on the MESA for the teacher and the distilled student are demonstrated in Table 5. The teacher shows strong and well-balanced performance across the hypnogram, achieving 85.71% accuracy, a macro-F1 of 0.8066, and Cohen's κ of 0.7778. It performs robustly on the dominant stages like Wake (F1 0.9135) and the merged Light class (F1 0.8612). The distilled student transfers this performance effectively on the well-represented stages, achieving 75.68% accuracy and retaining strong Wake (F1 0.8641) and Light (F1 0.7704) performance. Performance is more limited on the minority N3 and REM stages, where reduced recall lowers the student's F1 (N3 0.5264, REM 0.4218). This is due to the fact that these stages depend on sparse, multimodally encoded cues that are most sensitive to the capacity reduction imposed by distillation. We return to this in the teacher-ceiling analysis later in this section and treat it as an addressable limitation.

Table 5. Performance on the MESA cohort using the four-class scheme with N1 and N2 merged into a single Light-sleep class. Per-class precision, recall, and F1 scores with overall summary metrics are reported for the teacher and distilled student models as mean $\pm$ std over three seeds.

	Teacher			Student		
Stage	**P**	**R**	**F1**	**P**	**R**	**F1**
Wake	0.9409 ± 0.0314	0.8893 ± 0.0364	0.9135 ± 0.0043	0.8477 ± 0.0633	0.8855 ± 0.0412	0.8641 ± 0.0152
Light (N1 + N2)	0.8553 ± 0.0347	0.8689 ± 0.0384	0.8612 ± 0.0124	0.7590 ± 0.0341	0.7883 ± 0.0847	0.7704 ± 0.0235
N3	0.5851 ± 0.0704	0.7203 ± 0.0475	0.6429 ± 0.0364	0.5788 ± 0.0221	0.4888 ± 0.1036	0.5264 ± 0.0605
REM	0.8302 ± 0.0124	0.7894 ± 0.0523	0.8089 ± 0.0327	0.5491 ± 0.0932	0.3943 ± 0.2197	0.4218 ± 0.1552
Accuracy		0.8571 ± 0.0138			0.7568 ± 0.0183	
Macro-F1		0.8066 ± 0.0204			0.6457 ± 0.0286	
Weighted-F1		0.8592 ± 0.0127			0.7465 ± 0.0186	
Cohen's κ		0.7778 ± 0.0212			0.6134 ± 0.0243	

Table 6 reports the per-stage F1 of the teacher and distilled student on the MESA cohort under the five-class scheme, with the retention ratio (student F1/teacher F1). The student transfers the teacher's performance strongly on the stages that dominate the sleep period. Retention reaches 94.6% on Wake, 83.5% on N2, and 81.9% on N3, so the student preserves the large majority of the teacher's capacity on the most prevalent stages. Retention is lower on the minority N1 (76.3%) and REM (52.1%) stages. The contrast with the CFS cohort where retention remains high across most stages, indicates that distillation effectiveness is sensitive to how well the target cohort represents these sparse stages, an effect we analyse later in this section.

Table 6. Teacher–student per-stage F1 and retention on the MESA cohort (five-class), mean ± std over 3 seeds. Retention = (mean Student F1)/(mean Teacher F1).

Sleep Stage	Teacher F1	Student F1	Retention (%)
Wake (W)	0.9135 ± 0.0043	0.8641 ± 0.0153	94.6
N1	0.4969 ± 0.0373	0.3791 ± 0.0409	76.3
N2	0.7968 ± 0.0116	0.6654 ± 0.0292	83.5
N3	0.6429 ± 0.0364	0.5264 ± 0.0605	81.9
REM	0.8089 ± 0.0326	0.4218 ± 0.1552	52.1
Macro-F1	0.7318 ± 0.0212	0.5713 ± 0.0333	78.1
Cohen's κ	0.7147 ± 0.0153	0.5409 ± 0.0321	75.7

Table 7 presents the four-class staging results on the CFS cohort. The teacher achieves 82.00% accuracy, a macro-F1 of 0.8073, and Cohen's κ of 0.7336, with consistently strong and balanced performance across all four stages. The distilled student attained 76.55% accuracy, macro-F1 of 0.7454, and κ of 0.6556, maintaining strong per-stage performance including N3 (F1 0.7166) and the merged Light class (F1 0.7673), with a more noticeable reduction on REM.

Table 7. Performance on the CFS cohort using the four-class scheme with N1 and N2 merged into a single Light-sleep class. Per-class precision, recall, and F1 scores with overall summary metrics are reported for the teacher and distilled student models as mean ± std over three seeds.

	Teacher			Student		
Stage	**P**	**R**	**F1**	**P**	**R**	**F1**
Wake	0.8577 ± 0.0388	0.8726 ± 0.0329	0.8641 ± 0.0043	0.8425 ± 0.0252	0.8561 ± 0.0267	0.8488 ± 0.0084
Light (N1 + N2)	0.8600 ± 0.0157	0.7889 ± 0.0211	0.8228 ± 0.0174	0.8172 ± 0.0169	0.7261 ± 0.0676	0.7673 ± 0.0320
N3	0.6601 ± 0.0666	0.8184 ± 0.0544	0.7275 ± 0.0247	0.6842 ± 0.1235	0.7813 ± 0.1144	0.7166 ± 0.0268
REM	0.8074 ± 0.0649	0.8289 ± 0.0662	0.8147 ± 0.0197	0.5944 ± 0.0163	0.7164 ± 0.0421	0.6490 ± 0.0073
Accuracy		0.8200 ± 0.0133			0.7655 ± 0.0195	
Macro-F1		0.8073 ± 0.0127			0.7454 ± 0.0158	
Weighted-F1		0.8216 ± 0.0119			0.7672 ± 0.0199	
Cohen's κ		0.7336 ± 0.0187			0.6556 ± 0.0203	

The per-stage F1 of the teacher and distilled student on the CFS cohort under the five-class scheme is shown in Table 8. Retention is high across all four stages, ranging from 93.9% to 98.5%, and the student preserves 92.7% of the teacher's macro-F1 and 90.5% of its Cohen's κ. On the well-represented stages the student matches the teacher.

The stage with a meaningful loss is REM, where the student retains only 79.7% of the teacher's F1 ($0.8147 \rightarrow 0.6489$). Thus, even on CFS the student exhibits a REM-specific reduction, though it is considerably milder than on MESA. This is consistent with REM depending on temporally distributed multimodal cues that are the most sensitive to the capacity reduction imposed by distillation.

Additionally, in CFS cohort, a familygrouped split (grouping by family ID) in which no family appears in more than one partition is done. Under this protocol the teacher achieves macro-F1 0.7175 and Cohen's κ 0.7075, which is comparable to the split without family grouping (macro-F1 0.7050 ± 0.0098, Cohen's κ 0.6920 ± 0.0212 over 3 seeds). Since removing family overlap does not reduce performance, our results are not attributable to family leakage.

Table 8. Teacher–student per-stage F1 and retention on the CFS cohort (five-class), mean ± std over three seeds. Retention = (mean Student F1)/(mean Teacher F1).

Sleep Stage	Teacher F1	Student F1	Retention (%)
Wake (W)	0.8641 ± 0.0042	0.8488 ± 0.0085	98.2
N1	0.3281 ± 0.0134	0.3091 ± 0.0160	94.2
N2	0.7906 ± 0.0229	0.7423 ± 0.0378	93.9
N3	0.7275 ± 0.0244	0.7167 ± 0.0268	98.5
REM	0.8147 ± 0.0197	0.6489 ± 0.0073	79.7
Macro-F1	0.7050 ± 0.0098	0.6532 ± 0.0145	92.7
Cohen's κ	0.6920 ± 0.0212	0.6260 ± 0.0176	90.5

Table 9 illustrates the F1 score of teacher and student of both CFS and MESA cohorts. This table is intended to analyze and determine whether minority sleep stage degradation originates in distillation or in the teacher's representational limits. MESA and CFS teachers achieve comparable REM performance, 0.8089 and 0.8147 respectively. This shows that the teacher itself does not suffer a REM-specific deficit on the older MESA cohort. Despite this the student's REM performance degrades far more severely on MESA ($0.8089 \rightarrow 0.4218$) than on CFS ($0.8147 \rightarrow 0.6489$). The degradation is therefore concentrated in the distillation transfer rather than in the teacher's ceiling. The larger variance of the MESA student on REM $\pm$ 0.1552 versus $\pm$0.0073 on CFS indicates unstable learning of this stage under limited minority-class data. A similar but milder pattern is observed for N3, where the MESA student retains less of the teacher's F1 ($0.6429 \rightarrow 0.5264$) than the CFS student ($0.7275 \rightarrow 0.7167$), while majority stages such as Wake and N2 transfer robustly on both cohorts. These results localise the degradation to the distillation of sparse minority stages under limited data rather than to a ceiling imposed by the teacher.

Table 9. Teacher ceiling analysis: per-stage F1 (five-class) for teacher and distilled student on CFS and MESA cohorts, reported as mean ± std over three seeds.

Stage	CFS Teacher	CFS Student	MESA Teacher	MESA Student
Wake	0.8641 ± 0.0042	0.8488 ± 0.0085	0.9135 ± 0.0043	0.8641 ± 0.0153
N1	0.3281 ± 0.0134	0.3091 ± 0.0160	0.4969 ± 0.0373	0.3791 ± 0.0409
N2	0.7906 ± 0.0229	0.7423 ± 0.0378	0.7968 ± 0.0116	0.6654 ± 0.0292
N3	0.7275 ± 0.0244	0.7167 ± 0.0268	0.6429 ± 0.0364	0.5264 ± 0.0605
REM	0.8147 ± 0.0197	0.6489 ± 0.0073	0.8089 ± 0.0326	0.4218 ± 0.1552
Macro-F1	0.7050 ± 0.0098	0.6532 ± 0.0145	0.7318 ± 0.0212	0.5713 ± 0.0333
Cohen's κ	0.6920 ± 0.0212	0.6260 ± 0.0176	0.7147 ± 0.0153	0.5409 ± 0.0321

Table 10 compares the proposed framework against recent state-of-the-art methods on the MESA and CFS cohorts under the four-class scheme. To enable comparison with the existing literature, which predominantly reports four-class results, the proposed four-class metrics are obtained by merging the N1 and N2 stages of the five-class model. Baseline values are taken directly from the original publications and could not be re-evaluated under a common protocol. They differ from our models in input modality (PPG, cardiorespiratory, or video vs. full PSG), evaluation setting, and reported F1 variant, as annotated in the

table. Because these modalities and protocols differ, the comparison situates the proposed method within the broader literature rather than serving as a controlled benchmark. On both cohorts, the proposed teacher is competitive with specialized state-of-the-art methods. On MESA it attains a Cohen's κ of 0.7778, closely matching the leading PPG-based models SleepPPG-Net2 (κ 0.74) and InsightSleepNet (κ 0.742), and above the cardiorespiratory approach DCA-Sleep (κ 0.652). On CFS the teacher achieves κ 0.734 with 82.00% accuracy and a macro-F1 of 0.807, again competitive with the corresponding PPG-based results. The proposed models are trained and evaluated directly on each cohort, whereas SleepPPG-Net2 reports CFS as an out-of-distribution generalization target. The distilled student's four-class performance is lower, particularly on MESA, which is consistent with the minority-stage degradation discussed above. These results highlight that the primary advantage of the distilled student lies in its improved efficiency, while maintaining a favorable balance between computational efficiency and predictive performance.

Table 10. Comparison of four-class sleep-staging performance with existing methods on MESA and CFS. Competing methods use photoplethysmography (PPG), cardiorespiratory, or contactless video signals, whereas the proposed models use full PSG. Dashes denote values not reported in the source.

Model	Dataset	Modality	Accuracy (%)	Cohen's κ	Macro F1	Weighted F1
SleepPPG-Net2 [30]	MESA	PPG	84	0.74	–	–
InsightSleep [31]	MESA	PPG	84.2	0.742	–	0.842 [a]
SleepVST [56]	MESA	Video/cardioresp.	–	0.77	–	–
DCA-Sleep [57]	MESA	PPG	–	0.652	0.731 [b]	–
Proposed Model (T)	MESA	PSG	85.71	0.7778	0.8066	0.8592
Proposed Model (S)	MESA	PSG	75.68	0.6134	0.6457	0.7465
SleepPPG-Net2 [30] [c]	CFS	PPG	82	0.73	–	–
InsightSleep [31]	CFS	PPG	80.6	0.718	–	0.808
Proposed Model (T)	CFS	PSG	82.00	0.734	0.807	0.8216
Proposed Model (S)	CFS	PSG	76.55	0.6556	0.7454	0.7672

[a] base result before energy-score thresholding. [b] Macro/average F1 as reported by the source. [c] SleepPPG-Net2 evaluates CFS as an out-of-distribution generalization target; the proposed models are trained directly on each cohort.

Table 11 compares the five-class sleep-staging performance of the proposed models against three existing PSG-based methods on the MESA cohort. Among the baselines, performance varies considerably, with macro-F1 ranging from 0.5697 (PFTSleep) to 0.7610 (FullSleepNet), reflecting the inherent difficulty of five-class staging on this older, heterogeneous cohort. The proposed teacher achieves a macro-F1 of 0.7318 and Cohen's κ of 0.7147, exceeding both PFTSleep (0.5697) and the wavelet-based method (0.6820) by a substantial margin, and approaching FullSleepNet (0.7610). Notably, FullSleepNet is evaluated on a filtered subset in which 566 of the 2056 MESA recordings lacking five unique sleep stages or arousal labels are excluded; the proposed models are instead evaluated on the full retained cohort, which includes more heterogeneous and challenging recordings. The teacher's competitive performance under this less restrictive evaluation setting indicates that its accuracy is achieved without discarding difficult recordings. The distilled student achieves a macro-F1 of 0.5713 and κ of 0.5409, falling below the baseline methods while operating at a fraction of their computational cost (Table 12).

Table 11. Comparison of five-class sleep-staging performance with existing PSG-based methods on MESA. Competing methods are based on transformer, ensemble, and convolutional architectures, respectively. Baseline values are reported from the original publications. The proposed teacher (T) and distilled student (S) are trained and evaluated directly on the MESA cohort.

Model	Dataset	Modality	Accuracy (%)	Cohen's κ	Macro-F1
PFTSleep [58] [a]	MESA	PSG	69.57	0.560	0.5697
Wavelet-based Tech [59] [b]	MESA	PSG	83.80	0.627	0.6820
FullSleepNet [60] [c]	MESA	PSG	82.90	0.755	0.7610
Proposed Model (T)	MESA	PSG	79.35	0.7147	0.7318
Proposed Model (S)	MESA	PSG	66.28	0.5409	0.5713

[a] PFTSleep [58]: values as reported in the source. [b] Wavelet-based Tech [59]: values correspond to the combined unbalanced dataset as reported in the source (F1 = 68.2%, κ = 62.7%). [c] FullSleepNet [60] excludes 566 of the 2056 MESA recordings that lack five unique sleep stages or arousal labels, and is therefore evaluated on a filtered subset. The proposed models are evaluated on the full retained cohort, which includes more heterogeneous recordings.

These comparisons should also be interpreted in terms of the differing input modalities and their practical constraints. PPG and cardiorespiratory-based methods infer sleep stages indirectly from autonomic correlates such as pulse rate variability and respiratory effort rather than from direct neurophysiological activity. PPG arterial indices variability decreases as age increases even for healthy cohorts [61]. So PPG performance might degrade with factors common in the target population for home and long-term care monitoring. Cases of increasing age and obstructive sleep apnea severity alter the PPG waveform and weaken the autonomic signatures. Moreover, several of the strongest PPG-based methods require continuous full-night recordings (8–10 h) as input context to reach their reported performance, which limits their ability to provide timely, epoch-level feedback. PSG captures EEG, EOG, and EMG, providing richer and more direct discriminative information that depends on ocular and muscular activity not observable from PPG. The proposed framework is designed to exploit this multimodal information. In addition it focuses on deployable system so that the diagnostic richness of PSG is retained without the full computational and infrastructural burden of clinical systems. Table 12 compares the proposed distilled student against representative sleep-staging models in terms of parameter count. As these metrics are determined by model architecture and input configuration rather than by the evaluation dataset, the comparison is meaningful across methods trained on different cohorts. The distilled student contains only 0.168 M parameters, placing it firmly within the lightweight regime alongside dedicated edge-oriented models such as DistillSleep (0.109 M) and DistillSleepNet (0.080 M), and well below conventional architectures such as TinySleepNet (1.300 M), SleePyCo (1.630 M), and SleepTransformer (3.700 M). Crucially, this compact footprint is achieved while the proposed model processes full multimodal polysomnography (EEG, EOG, and EMG) through per-channel Conformer encoding, graph-based inter-channel modeling, and cross-attention fusion, whereas the comparison models operate on a single EEG channel. The proposed student therefore attains a parameter footprint comparable to single-channel edge models while supporting a substantially richer multimodal input.

Table 13 gives the computational profile of both teacher and student run on MESA cohort, benchmarked on an edge-class single-thread Intel CPU, per 30-s epoch, averaged over 100 timed forward passes after five warm-up runs. The student achieves gains in parameter count, where it falls from 6.468 M to 0.168 M (38.5 times reduction) and per-epoch floating-point operations from 1193.5 M to 144.0 M (8.3 times reduction). Measured under amortized streaming inference over 20-epoch windows, per-epoch latency drops from 74.94 ms to 6.64 ms, an 11.3× speed-up, while the fp32 parameter footprint shrinks from

24.69 MB to 0.64 MB (38.6× smaller). At 6.64 ms per epoch, a full 8-h recording (960 thirty second epochs) is scored in approximately 6 s, placing the distilled model well within the compute and memory budget of wearable and embedded sleep-monitoring hardware and enabling continuous on-device staging without cloud offload. Direct energy measurement requires instrumented hardware that is unavailable for this study. Given the 7 ms CPU latency, per-epoch energy consumption is correspondingly small, and hardware power profiling is identified as future work.

Table 12. Comparison of model complexity. Parameters are architecture-dependent and dataset-independent.

Model	Dataset	Params (M)
DistillSleepNet [26]	SleepEDF	0.080
DistillSleep [28]	SHHS	0.109
Micro SleepNet [29]	SleepEDF	0.048
ULW-SleepNet [62]	SleepEDF	0.013
SleePyCo [11]	SleepEDF	1.630
LEEGNet [39]	SleepEDF	0.970
TinySleepNet [27]	SleepEDF	1.300
SleepTransformer [63]	SleepEDF	3.700
Proposed (S)	CFS/MESA	0.168

Table 13. Computational profile of MESA Teacher vs. Student, measured on an edge-class CPU (Intel, single thread), per 30-s epoch, averaged over 100 forward passes.

Model	Params (M)	FLOPs (M)	Latency (ms/epoch)	Memory (MB)
Teacher	6.468	1193.5	74.94	24.69
Student	0.168	144.0	6.64	0.64
Compression	38.5×	8.3×	11.3×	38.6×

4.2. Discussion

This work set out to determine whether the diagnostic richness of full polysomnography could be retained in a compact model for on-device deployment in home-based and long-term care settings. The results across two clinical cohorts show that this is feasible, while also revealing that retention depends on how well minority sleep stages are represented in the target cohort.

The ablation on CFS shows that each design choice contributes to the final performance. Adding the EOG and EMG modalities via simple aggregation raises κ from 0.6457 to 0.6670 and accuracy from 74.01% to 76.01%, while macro-F1 dips slightly (0.6835 → 0.6711). Adding cross attention fusion increases all three metrics to the full-model values (77.51% accuracy, κ 0.6920, macro-F1 0.7050). The multimodal extension yields accuracy gain, confirming that ocular (EOG) and muscular (EMG) signals carry discriminative information that EEG alone does not recover.

On both cohorts the teacher attains agreement comparable to leading PPG-based methods, including on MESA, which comprises an older population. The proposed framework is trained and evaluated directly on each cohort, whereas the strongest comparators either treat one cohort as an out-of-distribution generalization target or depend on continuous full-night recordings for context. On a modality-matched, five-class comparison on MESA (Table 11), the teacher achieves a macro-F1 of 0.7318 and Cohen's κ of 0.7147, exceeding PFTSleep (0.5697) and a wavelet-based method (0.6820) and approaching FullSleepNet (0.7610). Notably, FullSleepNet is evaluated on a filtered subset that excludes 566 recordings lacking five unique sleep stages or arousal labels, whereas our models use the full retained cohort; the teacher's competitive performance under this less restrictive setting indicates

that its accuracy is not achieved by discarding difficult recordings. That other PSG methods also find five-class MESA staging challenging is consistent with the population-inherent difficulty of minority-stage discrimination in this older cohort. The teacher therefore matches specialized single-modality systems while offering epoch-level operation on a more informative but more complex input, indicating that it makes effective use of the multimodal signal.

The behaviour of the distilled student differs substantially between the two cohorts. On CFS, the student preserves 92.7% of the teacher's macro-F1 with strong per-stage transfer, although REM shows a meaningful reduction (79.7% retention). On MESA, the same student and identical distillation procedure transfer well on the dominant stages (94.6% retention on Wake, 83.5% on N2) but degrade substantially on some minority stages, REM (52.1% retention), N1 (76.3%) and N3 (81.9%). The fact that the same student transfers well on CFS but degrades on MESA indicates that minority-stage sparsity is the major contributing factor governing whether distillation succeeds. At the same time, the teacher retains strong REM performance on both cohorts (0.8147 on CFS, 0.8089 on MESA), showing that the sparse data is not inherently unlearnable, as confirmed by the teacher-ceiling analysis (Table 9). The limitation lies in transferring these sparse, multimodally-encoded stages to the compact student rather than in the data itself. The high across-seed variance of the MESA student on REM ($\pm$0.1552, versus $\pm$0.0073 on CFS) reinforces this interpretation, indicating unstable learning of this stage when minority examples are scarce. Because N3 and REM are both physiologically reduced and clinically significant in older adults, reliable performance on these stages is important. Since such sparsity is intrinsic to elderly populations and cannot be controlled at deployment, robustness to it is a design requirement that we will address in future work through minority-stage-aware distillation.

In terms of deployment, the distilled student occupies a footprint comparable to dedicated single-channel edge models such as DistillSleep (0.109 M) and DistillSleepNet (0.080 M) at 0.168 M parameters, yet uniquely processes information equivalent to full multimodal PSG through its EEG-only input and internal modality reconstruction. On an edge-class CPU it processes a 30-s epoch in 6.64 ms using 0.64 MB of memory an 11.3$\times$ latency and 38.6$\times$ memory reduction relative to the teacher confirming real-time, low-footprint operation, since per-epoch inference is far below the 30-s epoch duration. Because the deployed student requires only EEG, it also simplifies the sensing setup relative to multimodal systems, reducing the number of electrodes and the associated setup burden in care-facility and home environments.

5. Conclusions

This study presented a knowledge-distilled multimodal Conformer framework for on-device polysomnography-based sleep staging in home based and long-term care settings. The teacher model integrates per-channel Conformer encoders over CWT representation of EEG, a dynamic graph transformer for inter-channel dependency modeling, cross-attention fusion of EEG, EOG, and EMG, and a temporal transformer for inter-epoch context. To enable practical deployment, the framework distills this high-capacity teacher into a compact, EEG-only student that hallucinates the missing EOG and EMG representations, supervised through a combined feature-level and logit-level distillation objective. Evaluated on two clinical cohorts, CFS and MESA, the teacher achieves four-class staging performance competitive with modality-specialized state-of-the-art methods and an ablation confirms that both the multimodal extension and the cross-attention fusion contribute performance gains. The distilled student preserves 92.7% of the teacher's macro-F1 on CFS while requiring only 0.168 M parameters and running in 6.64 ms per 30-s epoch using 0.64 MB of memory on an edge-class CPU, confirming its suitability for real-time, low-power deploy-

ment. The distilled student largely preserves the teacher's minority-stage performance on the CFS cohort, but shows degradation on the sparser MESA cohort, particularly for REM and, to a lesser degree, for N1 and N3. This cohort difference suggests that the reduced minority-stage retention is primarily associated with limited minority-stage representation in MESA rather than a limitation of the teacher, which maintains strong performance across both cohorts. From a deployment perspective, the reduced retention on underrepresented stages highlights an important area for further improvement in the current method, particularly because minority-stage sparsity is inherent in elderly populations and cannot be directly controlled at inference time. We plan to address this through minority-stage-aware distillation and targeted supervision of the hallucination layer. Controlled disentanglement of contributing factors through cross dataset evaluation and matched-prevalence experiments forms a further direction for future work.

Author Contributions: Conceptualization, G.L.V.; methodology, S.S. and G.L.V.; software, G.L.V.; validation, S.S.; writing—original draft preparation, G.L.V.; writing—review and editing, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Mesa and CFS datasets used for this study are obtained from National Sleep Research Resource (NSRR). The accessibility to the data is obtained after a registration and approval of data request. The URLs of CFS, MESA are https://sleepdata.org/datasets/cfs (accessed on 18 May 2026), https://sleepdata.org/datasets/mesa (accessed on 18 May 2026).

Acknowledgments: The authors gratefully acknowledge Sri Mata Amritanandamayi Devi (Amma), Chancellor, Amrita Vishwa Vidyapeetham, for her inspiration and for providing financial support for the Article Processing Charges (APC) of this publication. The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by NIH-NHLBI Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups (RO1 HL098433). MESA is supported by NHLBI funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute, and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by NCATS. The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002).

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

References

1. Worley, S.L. The extraordinary importance of sleep: The detrimental effects of inadequate sleep on health and public safety drive an explosion of sleep research. *Pharm. Ther.* **2018**, *43*, 758.
2. Fernandes, M.; Antonucci, M.; Capecci, F.; Mercuri, N.B.; Della-Morte, D.; Liguori, C. Prevalence of sleep disorders in geriatrics: An exploratory study using sleep questionnaires. *Geriatr. Nurs.* **2024**, *60*, 107–113. [CrossRef] [PubMed]
3. Li, Y.; Tan, Y.; Zhao, Z. Impacts of aging on circadian rhythm and related sleep disorders. *Biosystems* **2024**, *236*, 105111. [CrossRef] [PubMed]
4. Ali, W.; Thalil, M. 0474 Insomnia Symptoms and Cognitive Frailty in Older Indians: Findings from a Nationally Representative Survey. *Sleep* **2025**, *48*, A207. [CrossRef]
5. Hoepel, S.J.W.; Leng, Y. Changes in sleep precede cognitive decline across two population cohorts: Unlocking the potential of sleep as an early marker of dementia. *Sleep* **2026**, *49*, zsaf268. [CrossRef] [PubMed]
6. Abhiram, A.; Rafeekh, A.; Surendran, S. Empathetic Companion: A Stateful, Agentic AI Architecture for Enhancing Safety and Alleviating Loneliness in Older Adults. *Procedia Comput. Sci.* **2026**, *282*, 1345–1356. [CrossRef]

7. Berry, R.B.; Brooks, R.; Gamaldo, C.; Harding, S.M.; Lloyd, R.M.; Quan, S.F.; Troester, M.T.; Vaughn, B.V. AASM Scoring Manual Updates for 2017, Version 2.4. *J. Clin. Sleep Med.* **2017**, *13*, 665–666. [CrossRef] [PubMed]
8. Patel, A.K.; Reddy, V.; Shumway, K.R.; Araujo, J.F. Physiology, sleep stages. In *StatPearls [Internet]*; StatPearls Publishing: St. Petersburg, FL, USA, 2024.
9. Supratak, A.; Dong, H.; Wu, C.; Guo, Y. DeepSleepNet: A Model for Automatic Sleep Stage Scoring Based on Raw Single-Channel EEG. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 1998–2008. [CrossRef] [PubMed]
10. Zhu, T.; Luo, W.; Yu, F. Convolution- and Attention-Based Neural Network for Automated Sleep Stage Classification. *Int. J. Environ. Res. Public Health* **2020**, *17*, 4152. [CrossRef] [PubMed]
11. Lee, S.; Yu, Y.; Back, S.; Seo, H.; Lee, K. SleePyCo: Automatic sleep scoring with feature pyramid and contrastive learning. *Expert Syst. Appl.* **2024**, *240*, 122551. [CrossRef]
12. Srinidhi, C.; Santhosh Kumar, C.; B, M.G.; Muralidharan, P.; Gopinath, S.; Anand Kumar, A. Improving the Performance of Sleep Apnea Screening System using Wavelets and Bottleneck Feature Fusion. In Proceedings of the 2022 International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Shivamogga, India, 14–15 October 2022; pp. 261–265. [CrossRef]
13. Harish Vijay, V.; Srichandra, I.V.; Kiran Patil, G.T.; Arunachalam, M.; Veluppal, A. Exploring Deep Learning Approaches for Sleep Apnea Detection. In *Proceedings of the Emerging Trends in Artificial Intelligence and Machine Learning*; Rai, J.K., Valada, A., Dubey, A.K., Abdelgawad, A., Eds.; Springer: Singapore, 2026; pp. 157–171.
14. Asmita, R.; Hrithik, T.; Akhil, K.H.; Mishra, N.; Kirthika Devi, V.S. Data-Driven Approaches for Personalized Sleep Disorder Diagnosis. In Proceedings of the 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 5–7 February 2025; pp. 124–129. [CrossRef]
15. Chung, W.; Hong, S.; Lee, W.; Bae, S. Transformer-Based Sleep Stage Classification Enhanced by Clinical Information. *arXiv* **2025**, arXiv:2511.08864.
16. Feng, Y.; An, Y.; Sun, H.; Li, Z.; Zhang, X. GateSleepNet: A Dual-Level Spatiotemporal Graph-Transformer Architecture for Automatic Sleep Staging. *IEEE Trans. Autom. Sci. Eng.* **2025**, *22*, 22860–22874. [CrossRef]
17. Kim, H.; Kim, M.; Chung, W. Sleep Staging Using Compressed Vision Transformer with Novel Two-Step Attention Weighted Sum. *IEEE Access* **2025**, *13*, 69650–69659. [CrossRef]
18. Geetha Lekshmy, V.; Anandu, P. Deep Learning for Sleep: An LSTM-Based Approach to Sleep Stage Prediction. In *Proceedings of the Computing and Machine Learning*; Springer: Singapore, 2026; pp. 557–566. [CrossRef]
19. Song, Y.; Zheng, Q.; Liu, B.; Gao, X. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2022**, *31*, 710–719. [CrossRef] [PubMed]
20. Zhao, W.; Jiang, X.; Zhang, B.; Xiao, S.; Weng, S. CTNet: A convolutional transformer network for EEG-based motor imagery classification. *Sci. Rep.* **2024**, *14*, 20237. [CrossRef] [PubMed]
21. S.S., P.; Menon, V.; Gopalan, S. Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition. *Biomed. Signal Process. Control* **2025**, *100*, 106967. [CrossRef]
22. Yubo, Z.; Yingying, L.; Bing, Z.; Lin, Z.; Lei, L. MMASleepNet: A multimodal attention network based on electrophysiological signals for automatic sleep staging. *Front. Neurosci.* **2022**, *16*, 973761. [CrossRef] [PubMed]
23. Phan, H.; Andreotti, F.; Cooray, N.; Chén, O.Y.; De Vos, M. Joint classification and prediction CNN framework for automatic sleep stage classification. *IEEE Trans. Biomed. Eng.* **2018**, *66*, 1285–1296. [CrossRef] [PubMed]
24. Cao, Y.; Xiang, W.; Wei, J.; Cao, S.; Tian, X.; Zhong, J.; Fang, X.; Luo, B.; Lyu, H.; Li, X. CrossFusionSleepNet: A multimodal deep learning model for automatic sleep stage classification. *Biomed. Signal Process. Control* **2026**, *112*, 108538. [CrossRef]
25. Shi, J.; Hao, Z.; Yue, H.; Chen, Z.; Liu, X.; Liu, X.; Li, J. Sleep stage classification with multi-resolution temporal modeling and adaptive feature fusion. *Biomed. Signal Process. Control* **2026**, *115*, 109224. [CrossRef]
26. Jia, Z.; Liang, H.; Liu, Y.; Wang, H.; Jiang, T. Distillsleepnet: Heterogeneous multi-level knowledge distillation via teacher assistant for sleep staging. *IEEE Trans. Big Data* **2024**, *11*, 1273–1284. [CrossRef]
27. 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), Montreal, QC, Canada, 20–24 July 2020; pp. 641–644. [CrossRef] [PubMed]
28. Park, K.; Hong, J.; Lee, W.; Shin, H.W.; Kim, H.S. DistillSleep: Real-time, on-device, interpretable sleep staging from single-channel electroencephalogram. *Sleep* **2025**, *48*, zsaf240. [CrossRef] [PubMed]
29. Liu, G.; Wei, G.; Sun, S.; Mao, D.; Zhang, J.; Zhao, D.; Tian, X.; Wang, X.; Chen, N. Micro SleepNet: Efficient deep learning model for mobile terminal real-time sleep staging. *Front. Neurosci.* **2023**, *17*, 1218072. [CrossRef] [PubMed]
30. Attia, S.; Shani Hershkovich, R.; Tabakhov, A.; Ang, A.; Oksenberg, A.; Tauman, R.; Behar, J.A. SleepPPG-Net2: Deep learning generalization for sleep staging from photoplethysmography. *Physiol. Meas.* **2025**, *46*, 125001. [CrossRef] [PubMed]
31. Nam, B.; Bark, B.; Lee, J.; Kim, I.Y. InsightSleepNet: The interpretable and uncertainty-aware deep learning network for sleep staging using continuous Photoplethysmography. *BMC Med. Inform. Decis. Mak.* **2024**, *24*, 50. [CrossRef] [PubMed]

32. Hinton, G.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. *arXiv* **2015**, arXiv:1503.02531. [CrossRef]
33. Jano, P.W. *Knowledge Distillation in Neural Networks: A Comprehensive Survey*; Technical Report; University of Wisconsin–Madison: Madison, WI, USA, 2025. [CrossRef]
34. Gou, J.; Sun, L.; Yu, B.; Wan, S.; Ou, W.; Yi, Z. Multilevel Attention-Based Sample Correlations for Knowledge Distillation. *IEEE Trans. Ind. Inform.* **2023**, *19*, 7099–7109. [CrossRef]
35. Chen, P.; Liu, S.; Zhao, H.; Jia, J. Distilling Knowledge via Knowledge Review. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 23–28 June 2021; pp. 5008–5017.
36. Jin, Y.; Wang, J.; Lin, D. Multi-Level Logit Distillation. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 24276–24285. [CrossRef]
37. Kim, H.; Kim, M.; Chung, W. Sleep-cmkd: Self-attention cnn/transformer cross-model knowledge distillation for automatic sleep staging. In *Proceedings of the 2023 11th International Winter Conference on Brain-Computer Interface (BCI)*; IEEE: New York, NY, USA, 2023; pp. 1–6.
38. Zhang, C.; Liao, Y.; Han, S.; Zhang, M.; Wang, Z.; Xie, X. Multichannel multidomain-based knowledge distillation algorithm for sleep staging with single-channel EEG. *IEEE Trans. Circuits Syst. II Express Briefs* **2022**, *69*, 4608–4612. [CrossRef]
39. Wang, B.; Li, Y.; Zhang, C.; Stefanidis, A.; Zhou, M.; Su, J. LEEGNet: Lightweight EEG Sleep Stage Classification Network with Knowledge Distillation. In Proceedings of the 2024 IEEE 5th International Conference on Pattern Recognition and Machine Learning (PRML), Chongqing, China, 19–21 July 2024; pp. 277–283. [CrossRef]
40. Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q.V. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–22 June 2019.
41. Ito, A.; Tanaka, T. SleepSatelightFTC: A Lightweight and Interpretable Deep Learning Model for Single-Channel EEG-Based Sleep Stage Classification. *IEEE Access* **2025**, *13*, 46263–46272. [CrossRef]
42. Yang, J.; Chen, Y.; Yu, T.; Zhang, Y. LMCSleepNet: A Lightweight Multi-Channel Sleep Staging Model Based on Wavelet Transform and Muli-Scale Convolutions. *Sensors* **2025**, *25*, 6065. [CrossRef] [PubMed]
43. Yang, C.; Li, B.; Li, Y.; He, Y.; Zhang, Y. LWSleepNet: A lightweight attention-based deep learning model for sleep staging with singlechannel EEG. *Digit. Health* **2023**, *9*, 20552076231188206. [CrossRef] [PubMed]
44. Xu, X.; Wang, Q.; Wang, C.; Zhang, Y. MCAF-Net: Multi-Channel Temporal Cross-Attention Network with Dynamic Gating for Sleep Stage Classification. *Sensors* **2025**, *25*, 4251. [CrossRef] [PubMed]
45. 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. [CrossRef] [PubMed]
46. Jia, Z.; Lin, Y.; Wang, J.; Zhou, R.; Ning, X.; He, Y.; Zhao, Y. GraphSleepNet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification. In *Proceedings of the IJCAI*; International Joint Conferences on Artificial Intelligence Organization: Yokohama, Japan, 2020; Volume 2021, pp. 1324–1330.
47. Wang, C.; Jiang, X.; Lv, C.; Meng, Q.; Zhao, P.; Yan, D.; Feng, C.; Xu, F.; Lu, S.; Jung, T.P.; et al. GraphSleepFormer: A multi-modal graph neural network for sleep staging in OSA patients. *J. Neural Eng.* **2025**, *22*, 026011. [CrossRef] [PubMed]
48. National Sleep Research Resource. Cleveland Family Study. Available online: https://sleepdata.org/datasets/cfs (accessed on 18 May 2026).
49. National Sleep Research Resource. Multi-Ethnic Study of Atherosclerosis (MESA). Available online: https://sleepdata.org/datasets/mesa (accessed on 18 May 2026).
50. Zhang, G.Q.; Cui, L.; Mueller, R.; Tao, S.; Kim, M.; Rueschman, M.; Mariani, S.; Mobley, D.; Redline, S. The National Sleep Research Resource: Towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 1351–1358. [CrossRef] [PubMed]
51. Chen, X.; Wang, R.; Zee, P.; Lutsey, P.L.; Javaheri, S.; Alcántara, C.; Jackson, C.L.; Williams, M.A.; Redline, S. Racial/ethnic differences in sleep disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep* **2015**, *38*, 877–888. [CrossRef] [PubMed]
52. Redline, S.; Tishler, P.V.; Tosteson, T.D.; Williamson, J.; Kump, K.; Browner, I.; Ferrette, V.; Krejci, P. The familial aggregation of obstructive sleep apnea. *Am. J. Respir. Crit. Care Med.* **1995**, *151*, 682–687. [CrossRef]
53. Sadowsky, J. Investigation of signal characteristics using the continuous wavelet transform. *Johns Hopkins Apl Tech. Dig.* **1996**, *17*, 258–269.
54. 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* **2020**, arXiv:2005.08100.
55. Peng, Z.; Huang, W.; Gu, S.; Xie, L.; Wang, Y.; Jiao, J.; Ye, Q. Conformer: Local features coupling global representations for visual recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 367–376.
56. Carter, J.F.; Jorge, J.; Gibson, O.; Tarassenko, L. SleepVST: Sleep Staging from Near-Infrared Video Signals using Pre-Trained Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 12479–12489.

57. Li, R.; He, Y.; Bie, Y.; Guo, J.; Wang, L.; Zhao, Y.; Zhong, J.; Zhu, W. Large-Scale Validation of a Dual Cross-Attention Network for Automated Sleep Staging Using Wearable Photoplethysmography Signals. *Diagnostics* **2026**, *16*, 802. [CrossRef] [PubMed]
58. Fox, B.; Jiang, J.; Wickramaratne, S.; Kovatch, P.; Suarez-Farinas, M.; Shah, N.A.; Parekh, A.; Nadkarni, G.N. A foundational transformer leveraging full night, multichannel sleep study data accurately classifies sleep stages. *Sleep* **2025**, *48*, zsaf061. [CrossRef] [PubMed]
59. Ingle, M.; Sharma, M.; Verma, S.; Sharma, N.; Bhurane, A.; Rajendra Acharya, U. Automated explainable wavelet-based sleep scoring system for a population suspected with insomnia, apnea and periodic leg movement. *Med. Eng. Phys.* **2024**, *130*, 104208. [CrossRef] [PubMed]
60. Zan, H.; Yildiz, A. Multi-task learning for arousal and sleep stage detection using fully convolutional networks. *J. Neural Eng.* **2023**, *20*, 056034. [CrossRef] [PubMed]
61. Yilmaz, G.; Ong, J.L.; Ling, L.H.; Chee, M.W. Insights into vascular physiology from sleep photoplethysmography. *Sleep* **2023**, *46*, zsad172. [CrossRef] [PubMed]
62. Wang, Z.; Zhou, D.; Xu, Q.; Cong, F.; Al-Sa'd, M.; Raitoharju, J. ULW-SleepNet: An Ultra-Lightweight Network for Multimodal Sleep Stage Scoring. In *Proceedings of the ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2026*; IEEE: New York, NY, USA, 2026; pp. 6596–6600.
63. Phan, H.; Mikkelsen, K.; Chén, O.Y.; Koch, P.; Mertins, A.; De Vos, M. Sleeptransformer: Automatic sleep staging with interpretability and uncertainty quantification. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 2456–2467. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.