Hierarchical Multimodal Sleep Representation Learning for Cross-Site Cognitive Impairment Prediction

Xinqi Bao and Saikat Chatterjee

KTH Royal Institute of Technology, Stockholm, Sweden

Abstract

Cognitive impairment progressively limits independent living, yet scalable methods for identifying individuals at future risk remain limited. In the 2026 George B. Moody PhysioNet Challenge, our team (Matcha) developed a hierarchical framework for predicting cognitive impairment from overnight polysomnography (PSG). Modality-specific convolutional stems learned a 192-dimensional representation for every 30-s epoch from human and automated sleep annotations together with cross-view self-supervision. Local and full-night Transformers then aggregated the epoch sequence into a patient-level prediction, while a low-capacity record-wise residual represented whole-night sleep summaries and outcome-observation timing. In the three-site leave-one-site-out (LOSO) validation, it achieved a macro age-conditioned AUROC of 0.804. On the official phase hidden validation set, the best Small- and Large-track entries achieved 0.847 and 0.844, respectively, ranking 1st out of 513 entries.

1. Introduction

Cognitive impairment affects memory, reasoning, and independent living, but early population-level screening remains difficult. Brief cognitive tests can miss subtle impairment [1], whereas amyloid, tau, and neurodegeneration biomarkers are often invasive or costly for broad screening [2]. Sleep offers a complementary physiological window into brain health. Sleep-disordered breathing and hypoxia, for example, have been associated with subsequent mild cognitive impairment and dementia [3]. Polysomnography (PSG) simultaneously measures neural, cardiac, respiratory, oxygenation, and motor activity and may therefore support opportunistic cognitive risk assessment during routine sleep studies.

Most automated PSG research has focused on sleep staging and sleep disorders. Recent large-scale work showed that multimodal PSG representations can also predict future diseases, including dementia [4]. Translating this idea to cognitive impairment nevertheless remains challenging. A single night contains approximately 800–1,200 epochs, while the outcome is available only at the patient level years later. Recordings also differ in channel montage, sensors, filtering, and missing modalities across institutions. A successful model must preserve local physiological and event information, summarize its organization across the night, and avoid relying solely on acquisition-site signatures.

The 2026 George B. Moody PhysioNet Challenge provided a curated subset of the Human Sleep Project for predicting a cognitive impairment diagnosis one to six years after a sleep study [5]. The proposed framework has three components: a multimodal 30-s encoder trained with human, automated, and self-supervised targets; local and full-night Transformers trained for patient-level prediction; and a low-capacity record-wise residual for sleep summaries and outcome-observation timing. Site-wise validation and official hidden-site evaluation were used to assess cross-institutional generalization.

2. Methodology

2.1. Data and preprocessing

The Challenge data were drawn from the Human Sleep Project [6]. Source codes identify three deidentified institutional cohorts. The Large set contained 6,600 records from 6,600 patients: S0001 ($n = 5,139$), I0006 ($n = 1,142$), and I0002 ($n = 319$), with 498 positive outcomes. The prevalence-matched Small set contained 1,103 patients: S0001 ($n = 857$), I0006 ($n = 192$), and I0002 ($n = 54$), with 84 positives.

Available signals were mapped to 20 canonical channels: six electroencephalography (EEG), two electrooculography (EOG), one electrocardiography (ECG), seven respiratory, one peripheral oxygen saturation (SpO_2), and three electromyography (EMG) channels. EEG, EOG, ECG, and EMG were filtered and resampled to 128 Hz, respiration to 32 Hz, and SpO_2 to 1 Hz. Each channel was robustly centered and scaled within the night. Five-second quality masks marked flat, invalid, or extreme segments. Channel availability was retained explicitly, and missing

channels were zero-filled only after masking. Valid signals were divided into 30-s epochs.

2.2. Multimodal epoch representation

Thirty-second windows align with sleep staging and provide a tractable unit for learning local rhythms, morphology, and short events before patient-level aggregation. As shown in Fig. 1(a), each of the six modalities was processed by a separate one-dimensional convolutional stem containing three strided convolutions and a dilated residual block. Each stem produced eight temporal tokens. Attention pooling yielded six modality vectors (248 values), which were concatenated with 20 channel-availability indicators and mapped by a 268–320–192 multilayer perceptron to a 192-D epoch embedding.

The encoder combined strong, weak, and self-supervision. Five-class sleep staging and arousal, respiratory-event, and limb-event heads used all available human labels and Complete Artificial Intelligence Sleep Report (CAISR) targets [7]; unavailable labels were masked. Human stage, CAISR stage, human event, and CAISR event losses received weights of 1.0, 0.5, 0.4, and 0.25, respectively. This supervision encouraged the embedding to retain established sleep semantics rather than only a coarse patient summary.

For self-supervision, two independently perturbed views were generated from the same 30-s epoch. A student encoded one view and was trained to match global and modality-specific targets produced from the other view by an exponential-moving-average teacher. Thus, acquisition-style changes could not alter the physiological identity assigned to the epoch. This followed consistency-teacher and bootstrapped latent-prediction principles [8,9]. The fixed augmentation and sampling settings are summarized in Table 1. The resulting exponential-moving-average teacher generated 5,895,218 eligible 192-D embeddings and was frozen for cognitive impairment (CI) training.

Table 1. Encoder training and augmentation settings.

Signal style	Gain/offset, low-/high-frequency mixing, Gaussian noise, and quantization
Occlusion	Temporal masking, channel dropout (10%), and whole-modality dropout (8%)
Training	15 epochs; 128 windows per night in each epoch

Encoder behavior was assessed on a fixed 600-night pilot. Diagnostics included human sleep-stage recall; the mean Spearman correlation of linear probes with representative EEG spectral, heart-rate-variability, respiratory, SpO_2, and EMG measurements; cosine similarity before and after channel masking; the corresponding loss of stage recall; and linear site-classification accuracy. These measures quantified retained physiology, missing-channel stability, and residual acquisition-site information.

2.3. Hierarchical whole-night model

Figure 1(b) summarizes the hierarchical whole-night model. The frozen epoch sequence was projected from 192 to 256 dimensions. A two-layer, eight-head local Transformer [10] processed non-overlapping groups of ten epochs, followed by masked mean pooling to form one 5-min token. This stage had no direct target: its representation was learned through the patient-level CI loss. A learned class token and positional embeddings were then added to the 5-min sequence, and a three-layer, eight-head full-night Transformer produced a 256-D night representation and CI logit.

Weighted binary cross-entropy (BCE) compensated for 498 positive and 6,102 negative patients. To align training with the official age-conditioned AUROC, the objective also included a pairwise logistic loss with weight 0.15. This term ranked a positive patient above a negative patient from the same site when their ages differed by at most two years. Four natural batches were followed by one site-balanced batch. Checkpoints were selected only by age-conditioned AUROC. Three independently initialized models were ensembled; their training durations of one to four epochs were selected by inner validation.

2.4. Record-wise residual and deployment

The outcome definition depends on both diagnosis timing and available follow-up. Three low-capacity corrections were therefore fitted on training data: a standardized recording-date coefficient; a six-year follow-up eligibility score derived from each training site's 99th-percentile last visit; and an L2-regularized linear model over 90 whole-night CAISR summary features. All coefficients optimized legal within-site, age-matched pairs. The deployed score added 0.375 times the sum of these corrections to the PSG ensemble logit. Inference remained record-wise: no hidden-cohort statistic was accessed, missing components caused zero adjustment, a missing timestamp could be read from the current EDF header, and PSGs shorter than one complete epoch used a finite fallback.

3. Results

Table 2 reports CI-label leave-one-site-out (LOSO) validation on the Large set. Here, the held-out site's CI labels were excluded from CI training, although the encoder had previously seen auxiliary PSG supervision from all three training sites. Simple pooled embedding statistics reached 0.649 macro age-conditioned AUROC. Hierarchical modeling improved every site, and the three-seed PSG

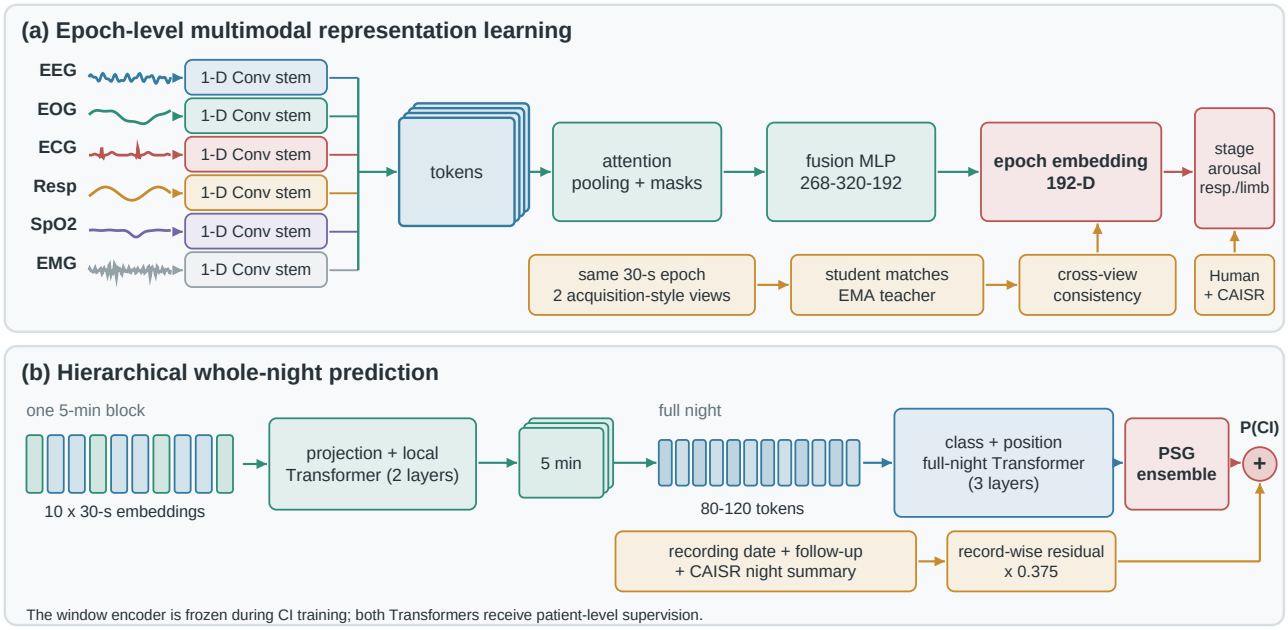

Figure 1. Proposed framework. (a) Six modality-specific stems encode each 30-s epoch. Human and CAISR targets provide sleep-stage and event supervision, while two perturbed views of the same epoch provide a consistency target. (b) Groups of ten epoch embeddings form 5-min tokens, which a full-night Transformer maps to a patient-level CI score. A low-capacity record-wise residual is added to the three-seed PSG sequence ensemble.

sequence ensemble reached 0.727. The record-wise residual increased macro performance to 0.804.

Table 2. Site-wise age-conditioned AUROC on the Large training set.

Model	I0002	I0006	S0001	Macro
Pooled epoch statistics	0.671	0.688	0.587	0.649
Hierarchical PSG, one seed	0.747	0.755	0.642	0.715
Hierarchical PSG, three seeds	0.761	0.752	0.669	0.727
+ record-wise residual	**0.918**	**0.784**	**0.711**	**0.804**

In a matched outer-fold replay, recording date and follow-up eligibility increased macro AUROC from 0.730 to 0.803; CAISR summaries raised it to 0.804. Age, sex, body mass index, and session identifiers improved local LOSO but reduced official I0004 age-conditioned AUROC from 0.844 to 0.828 and were therefore excluded.

Device-style consistency was examined independently on a fixed 600-night pilot (Fig. 2). Relative to an otherwise identical encoder, it preserved human stage recall, improved recovery of representative physiological measurements, and substantially increased missing-channel stability, while site-classification accuracy decreased modestly. These results showed that the consistency objective reduced acquisition sensitivity without sacrificing the supervised sleep-stage task.

On official hidden validation, the best Small- and Large-track entries achieved 0.847 and 0.844 age-conditioned

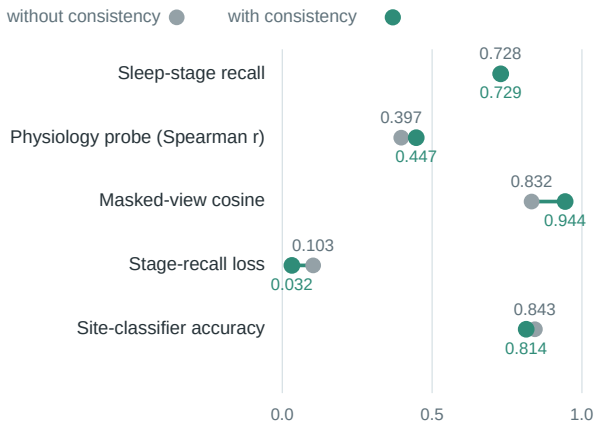

Figure 2. Encoder diagnostics without (gray) and with (green) cross-view consistency. The physiology probe is the mean Spearman correlation for the measurements described in Section 2.2. Higher is better for the first three rows; lower is better for the final two.

AUROC, placing team Matcha 1st out of 513 scored entries. The selected, fully server-retrained Large entry reran the encoder, embedding, sequence, and residual stages from the provided training set on the official server. It achieved 0.840 age-conditioned AUROC with

a prevalence-based reward of 0.336 (Table 3).

Table 3. Official hidden I0004 validation results.

Entry	AC-AUROC	AUROC	AUPRC	Reward
Best (Small)	**0.847**	**0.879**	**0.558**	0.286
Best (Large)	0.844	0.878	0.528	0.291
Selected (Large)	0.840	0.878	0.500	0.336

4. Discussion and Conclusion

The PSG pathway contributed at two levels. First, pooled 30-s embeddings already contained patient-level CI information. Second, hierarchical modeling raised macro AUROC from 0.649 to 0.727, with gains at all three held-out training sites. Because the hierarchy also changes model capacity and nonlinearity, this gain cannot yet be attributed exclusively to temporal order; shuffled-order controls remain future work.

Recording chronology and follow-up eligibility supplied the largest residual gain. This is plausible because a negative outcome requires six subsequent years without a CI diagnosis, whereas a positive outcome requires diagnosis within the same horizon. The residual therefore models label observability as well as patient physiology and should not be interpreted as a biomarker. Its transfer risk was limited by training-only normalization, record-wise inference, zero adjustment for missing fields, and preservation of the strong PSG pathway.

The internal evaluation was CI-label LOSO rather than strict representation-domain LOSO because encoder pretraining used auxiliary PSG targets from all three training sites. The hidden validation cohort was the first acquisition site unseen by both encoder and CI model. Strong transfer to that cohort is encouraging, but a separate hidden test site may differ in population, hardware, channels, and metadata availability. Other limitations include only three training sources, 498 positive patients, incomplete human annotations, and possible systematic errors in algorithmic targets.

In conclusion, multimodal PSG pretraining, hierarchical whole-night aggregation, metric-aligned CI learning, and a low-capacity record-wise residual yielded substantial cross-site improvement. Team Matcha ranked 1st out of 513 entries on the official validation leaderboard, achieving age-conditioned AUROC of 0.847 and 0.844 on the Small and Large tracks. These findings support task-specific multimodal representation learning and hierarchical whole-night modeling for cross-site cognitive impairment screening.

References

[1] Nasreddine ZS, Phillips NA, Bédirian V, et al. The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. Journal of the American Geriatrics Society 2005;53(4):695–699.

[2] Jack CR, Bennett DA, Blennow K, et al. NIA-AA research framework: Toward a biological definition of Alzheimer's disease. Alzheimers Dementia 2018;14(4):535–562.

[3] Yaffe K, Laffan AM, Harrison SL, et al. Sleep-disordered breathing, hypoxia, and risk of mild cognitive impairment and dementia in older women. JAMA 2011;306(6):613–619.

[4] Thapa R, Kjaer MR, He B, et al. A multimodal sleep foundation model for disease prediction. Nature Medicine 2026; 32(2):752–762.

[5] Reyna MA, Weigle AT, Li Q, et al. Screening for Cognitive Impairment During Sleep Studies: The George B. Moody PhysioNet Challenge 2026. In Computing in Cardiology, volume 53. 2026; In press.

[6] Sun H, Ganglberger W, Nasiri S, et al. The Human Sleep Project (version 2.0), 2023. Brain Data Science Platform. doi:10.60508/qjbv-hg78.

[7] Nasiri S, Ganglberger W, Nassi T, et al. CAISR: Achieving human-level performance in automated sleep analysis across all clinical sleep metrics. Sleep 2025;48(8):zsaf134.

[8] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In Advances in Neural Information Processing Systems, volume 30. 2017; 1195–1204.

[9] Grill JB, Strub F, Altche F, et al. Bootstrap your own latent: A new approach to self-supervised learning. In Advances in Neural Information Processing Systems, volume 33. 2020; 21271–21284.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. 2017; 5998–6008.

Address for correspondence:

Xinqi Bao, KTH Royal Institute of Technology, Stockholm, Sweden

xba@kth.se