Deep Learning of Multimodal Features Derived from Raw Polysomnography for Cognitive Impairment Prediction

Somesh Saini[1], Sikawat Thanaviratananich[2], Matheus Araujo[2,3]

[1] Computer and Data Sciences, Case Western Reserve University, Cleveland, USA
[2] Sleep Disorders Center, Cleveland Clinic Foundation, Cleveland, USA
[3] Biomedical Engineering, Case Western Reserve University, Cleveland, USA

Abstract

Cognitive impairment is often diagnosed years after symptoms begin, motivating earlier, noninvasive risk assessment. We investigated whether multimodal features derived from overnight polysomnography can predict cognitive impairment 1-6 years after a sleep study. Each 30-second epoch was represented by 105 features, including EEG, EMG, SpO_2, heart-rate variability, and automated sleep annotations. We then fused the epoch-level features with whole-night aggregated features and demographics using a temporal convolutional network for binary prediction of cognitive impairment. As part of the George B. Moody PhysioNet Challenge 2026, our team (CLECLINIC) achieved an age-conditioned AUROC of 0.696 and a prevalence-weighted reward of 0.125 on the official validation set. These results support the potential of routine polysomnography for opportunistic cognitive-risk assessment.

1. Introduction

Cognitive impairment (CI) is a clinical spectrum characterized by difficulties in thinking, learning, remembering, or making decisions. As cognitive deficits progress, they can increasingly interfere with functional independence and, in some individuals, progress into dementia. Dementia, including Alzheimer's disease, vascular dementia, and dementia with Lewy bodies, is the most severe stage of cognitive impairment and affects an estimated 57 million people worldwide, a number projected to nearly triple to 153 million by 2050 [1]. Despite this growing burden, diagnosis is often delayed, with symptoms preceding formal diagnosis by an average of 3.5 years [2]. Existing approaches to early detection face a tradeoff between accessible but nonspecific cognitive screening (Montreal Cognitive Assessment) and more definitive but costly and specialist-dependent biomarker testing—including cerebrospinal fluid analysis and neuroimaging such as amyloid positron emission tomography—motivating the need for accessible, noninvasive approaches that can identify CI earlier [3, 4].

Polysomnography (PSG) offers a potential source of such information by capturing a range of physiological signals, including electroencephalography (EEG), electromyography (EMG), electrocardiography (ECG), respiratory effort and airflow, and blood oxygen saturation. These signals provide measures of sleep architecture and physiology that have been associated with subsequent cognitive decline [5]. Specifically, reduced rapid eye movement (REM) sleep and longer REM latency have been associated with incident dementia over follow-up periods of up to 19 years [6], while longitudinal reductions in slow-wave sleep have also been associated with increased dementia risk [7]. Together, these findings suggest that PSG contains information relevant to cognitive decline beyond any single sleep measure, providing a basis for automated, noninvasive risk assessment.

The George B. Moody PhysioNet Challenge 2026 builds on these findings by proposing a prediction task: given a single PSG, a model should predict whether a patient will receive a diagnosis of CI 1-6 years later [8]. As part of the Challenge, our team, CLECLINIC, developed a deep learning approach that leverages multimodal physiological features derived from polysomnography to address this prediction task. This work presents our proposed framework and evaluation.

2. Methods

The following subsections present the model architecture and evaluation approach for CI prediction.

2.1. Data

The training set contains 6,600 patients (62.0 ± 8.5 years; 53.0% male, n=3,501; 47.0% female, n=3,099), with one PSG session per patient. The data originate from the Human Sleep Project (HSP) [9] and were adapted for the George B. Moody PhysioNet Challenge 2026 [8]. In the cleaned, multi-site training dataset, CI prevalence is 7.55% with 498 positive and 6,102 negative cases. We applied a

smaller 200 GB subset for internal cross-validation across 780 patients (70.4 $\pm$ 8.3 years old; 60.4% male).

2.2. Feature representation

Traditional sleep analysis often relies on whole-night measures of sleep macroarchitecture, such as the total time spent in stages N1, N2, N3, and REM sleep. Although these measures provide useful summaries of overall sleep structure, they may overlook finer-grained physiological patterns that are also associated with cognitive impairment. We therefore construct epoch-level representations from clinically motivated features derived from EEG, EMG, SpO2, ECG, and sleep and event information derived from the Complete AI Sleep Report (CAISR) [10].

Each 30-second epoch is represented by 105 features, including 90 physiological features, a 5-dimensional CAISR sleep-stage encoding, 7 validity indicators, and 3 time-of-night position channels. The physiological representation contains 72 EEG features across six bipolar derivations, with each derivation contributing relative power in six frequency bands, a slowing ratio, an alpha/theta ratio, Hjorth mobility and complexity [11], waveform kurtosis, and line length (12 features per derivation). The latter two are time-domain components of the per-epoch feature set behind the Brain Age Index of Sun et al. [12], which we carry in raw per-epoch form. Here, the index itself is a whole-night quantity and is handled in the static vector below. The remaining 18 physiological features summarize chin (3) and leg (2) EMG, SpO2 desaturation (2), and spindle and slow-oscillation microstructure (6), alongside CAISR respiratory, arousal, and limb-movement events (3) [13]. Additionally, 2 heart-rate variability features were derived using Pan-Tompkins R-peak detection [14].

A separate 33-dimensional static vector contains whole-night demographics, sleep macroarchitecture, SpO2 measures, including hypoxic burden [15], whole-night spindle and slow-wave densities, CAISR event indices, heart rate variability summaries and dynamics, and the brain age gap. This brain-gap is the bias-corrected residual of a ridge regression fit out-of-fold on cognitively normal training subjects, which predicts chronological age from each patient's night-aggregated EEG microstructure [16].

2.3. Preprocessing

PSG recordings differ across sites in channel naming, sampling rate, amplifier gain, and channel availability [17]. Therefore, we resample signals to 200 Hz and cut into 30-second epochs, capped at 1024 epochs per recording, with labeling for the six EEG channels, chin and leg EMG, SpO2, and ECG. When a bipolar EEG channel is unavailable, it is reconstructed by subtracting the reference from the exploring electrode on a sample-by-sample basis, so that C4-M1 is obtained as C4 subtracted from M1. Because absolute amplitude is not comparable across amplifiers, amplitude-sensitive features are normalized within subject by the median and interquartile range [18]. Highly skewed power features are compressed with a $\log(1 + |x|)$ transform. During training, whole modalities are randomly blanked to expose the model to missing-modality conditions.

Table 1: Parameter budget of model implementation.

Stage	Configuration	Params
Input	105 features $\times$ T epochs	—
Cross-family attention		
Family tokenizers	8$\times$ Linear	**3,296**
Self-attention	4 heads	**4,224**
Layer normalization	$d = 32$	**64**
Family projections	8$\times$ Linear	**3,135**
Residual gates	8 scalars, init. 0.1	**8**
InceptionTime blocks		
Block 1	$105 \to 128$ channels	**243,392**
Blocks 2-4	$128 \to 128$; 247,808 each	**743,424**
Fusion head		
Hidden layer 1	Linear($929 \to 256$), ReLU	**238,080**
Hidden layer 2	Linear($256 \to 128$), ReLU	**32,896**
Output	Linear($128 \to 1$)	**129**
One network		**1,268,648**
Ensemble	3 networks, averaged	**3,805,944**

2.4. Backbone

We apply a deep learning approach to integrate extracted clinical features for predictive capabilities. We precede the temporal blocks with one round of cross-family attention. The family modality includes EEG, chin/leg EMG, SpO2, spindle microstructure, CAISR events, heart-rate variability, and sleep stage. This is projected into a shared 32-dimensional token and four-head self-attention that runs across the eight tokens within each epoch [19, 20].

The attended sequence passes to InceptionTime blocks, where each holds three Inception modules [21]. A module applies a 1×1 convolution of width 32, then three parallel convolutions with kernel sizes 9, 19, and 39 epochs, then a max-pool branch. Per epoch, these kernels span 4.5, 9.5, and 19.5 minutes, matching the timescale of sleep-stage cycling rather than any raw-signal frequency. By pooling by stage, we identify sleep-localized features, such as reduced REM sleep and reduced slow-wave sleep [7].

The pooled slots are passed to a two-layer head of 256 and 128 units with ReLU and dropout. We train three networks differing only in random initialization and average their sigmoid outputs [21]. The model is implemented in Python 3.10 and PyTorch 2.2.0. One network holds 1.27M parameters, and the three-network ensemble holds 3.81M, as in Table 1. Table 2 gives the search space and the deployed values.

2.5. Evaluation

The Challenge uses two age-aware evaluation metrics. The primary ranking metric is the age-conditioned area under the receiver operating characteristic curve (AUROC), which measures how well the model ranks patients with

Table 2: Hyperparameter search space(with Optuna), maximizing age-conditioned AUROC on 200GB data subset.

Hyperparameter	Search space	Final
Inception blocks	{2, 3, 4}	4
Filters per branch	{24, 32, 48, 64}	32
Head dropout	[0.2, 0.6]	0.51
Pooled dropout	[0.0, 0.5]	0.23
Batch size	{16, 32, 64, 124}	64
Learning rate	[3e-4, 3e-3]	2.7e-3
Weight decay	[1e-5, 1e-3]	4.3e-4
Rank-loss weight	[0, 0.6]	0.43

Table 3: Official results scored by PhysioNet Challenge.

Metric	Official validation
Age-conditioned AUROC	0.696
Prevalence-weighted reward	0.125
Age-weighted AUROC	0.679
AUROC	0.758
AUPRC	0.185
Accuracy	0.714
F-measure	0.213

cognitive impairment above patients without cognitive impairment of similar age. The prevalence-based reward metric evaluates binary predictions while accounting for the age-specific prevalence of cognitive impairment. Both metrics compare patients within ± 2 years of age to reduce the influence of age on model performance [8].

To be consistent with the PhysioNet Challenge framework, we submitted all trained models to the official challenge organizers. The submitted model performance was evaluated on a hidden validation and test set. Since the Challenge setting included multiple competitors and limited each team to only ten submission attempts, careful model selection and tuning were necessary.

2.6. Interpretability and Ablation Analysis

We performed complementary ablation and Shapley additive explanation (SHAP) analyses to assess the contribution of the physiological feature families to model predictions [22]. For ablation, EEG, EMG, ECG-derived HRV, SpO_2, and CAISR features were independently removed, and the model was retrained from scratch under patient-grouped 3-fold cross-validation. We also computed family-level SHAP values to quantify the direction and magnitude of each clinical feature family's contribution to the model output.

3. Results

On the official validation set, our system reached an age-conditioned AUROC of 0.696 at a reward of 0.125.

Additional analyses examined the contribution of individual feature families to model predictions. As shown in Figure 1, family-level SHAP values revealed the largest and most variable contributions from event-related features, followed by sleep architecture, HRV, and EEG, whereas biomarkers, demographics, SpO_2, and EMG showed comparatively smaller effects. The missing-

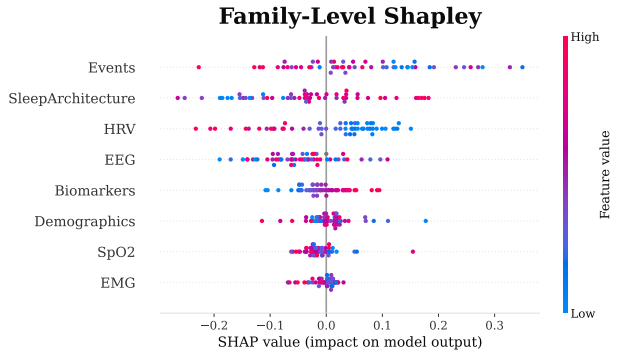

Figure 1: Family-level Shapley beeswarm across the nine clinical feature families. Each point is one patient.

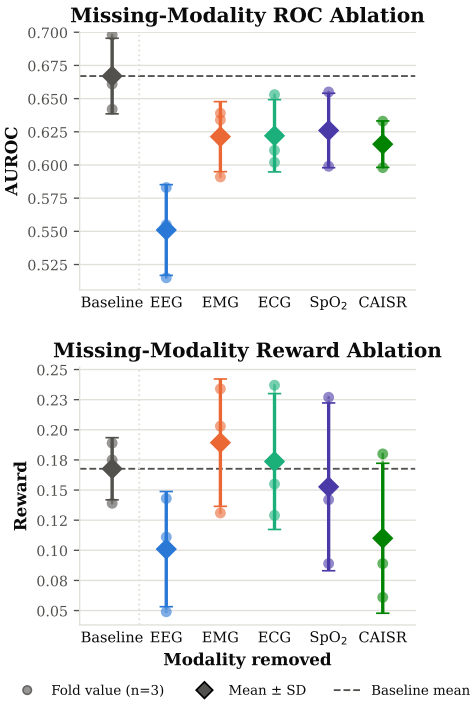

Figure 2: Missing-modality ablation under patient-grouped 3-fold cross-validation. Each condition retrains with one signal family removed.

modality analysis in Figure 2 further demonstrated that EEG provided the strongest discriminative contribution: removing EEG reduced the mean age-conditioned AUROC from 0.670 to 0.551 across the three cross-validation folds. Removal of the remaining modalities produced substantially smaller changes in AUROC, although larger variation was observed for the prevalence-weighted reward, particularly following removal of EEG and CAISR-derived features.

4. Discussion

Our system achieved an age-conditioned AUROC of 0.696 and a prevalence-weighted reward of 0.125 on the official validation set. Together, these results provide evidence for the general direction of using clinically derived, multimodal PSG features to identify information associated with future cognitive impairment. The pronounced EEG-removal drop seen in Figure 2) is consistent with EEG-derived sleep architecture's established link to dementia risk, suggesting the model relies primarily on EEG-derived signal [6, 7].

Events in the PSG exhibited the largest overall contribution according to the SHAP values analysis. Sleep architecture and HRV also showed notable contributions, with higher sleep-architecture values tending toward positive SHAP values and lower HRV values tending toward positive contributions. EEG showed a substantial but more heterogeneous distribution of contributions, while other features had comparatively smaller overall effects.

Several limitations should be considered. Requiring at least six years of subsequent clinical encounters for negative cases may preferentially include patients with sustained healthcare access, limiting generalizability to populations with less consistent follow-up. Additionally, the PSG data span multiple clinical sites, with differences in equipment and channel availability that may influence predictions despite modality masking during training.

Future work should evaluate the approach on independently collected cohorts while assessing robustness across healthcare systems and PSG acquisition protocols.

5. Conclusion

We demonstrate that clinically derived features from a single overnight PSG can provide predictive information about future cognitive impairment. Our approach achieved an age-conditioned AUROC of 0.696 and a prevalence-weighted reward of 0.125 on the official validation set, supporting the potential of routine sleep recordings for opportunistic cognitive risk assessment. Further validation across independent cohorts and clinical settings is needed to establish its generalizability and clinical utility.

References

[1] E. Nichols, J. D. Steinmetz, S. E. Vollset *et al.*, "Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease study 2019," *The Lancet Public Health*, vol. 7, no. 2, pp. e105–e125, 2022.

[2] O. Kusoro, M. Roche, R. Del-Pino-Casado, P. Leung, and V. Orgeta, "Time to diagnosis in dementia: A systematic review with meta-analysis," *International Journal of Geriatric Psychiatry*, vol. 40, no. 7, p. e70129, 2025.

[3] D. H. J. Davis, S. T. Creavin, J. L. Y. Yip, A. H. Noel-Storr, C. Brayne, and S. Cullum, "Montreal cognitive assessment for the diagnosis of Alzheimer's disease and other dementias," *Cochrane Database of Systematic Reviews*, no. 10, p. CD010775, 2015.

[4] A. Leuzy *et al.*, "Considerations in the clinical use of amyloid PET and CSF biomarkers for Alzheimer's disease," *Alzheimer's & Dementia*, vol. 21, no. 3, p. e14528, 2025.

[5] E. Bilal, M. L. D. Araujo, K. L. Beck, C. M. Heinzinger, S. Ghosn, C. Y. Saab, N. Foldvary-Schaefer, J. L. Rogers, and R. Mehra, "A foundation model for sleep-based risk stratification and clinical outcomes," *Nature Communications*, vol. 17, no. 1, p. 7603, 2026.

[6] M. P. Pase, J. J. Himali, N. A. Grima *et al.*, "Sleep architecture and the risk of incident dementia in the community," *Neurology*, vol. 89, no. 12, pp. 1244–1250, 2017.

[7] J. J. Himali, A.-A. Baril, M. G. Cavuoto *et al.*, "Association between slow-wave sleep loss and incident dementia," *JAMA Neurology*, vol. 80, no. 12, pp. 1326–1333, 2023.

[8] M. A. Reyna, A. Weigle, Q. Li, Z. Koscova, H. Sun, S. Wen, U. Katwa, R. Thomas, D. Hwang, L. M. Trotti, E. Mignot, R. Sameni, S. Nasiri, M. B. Westover, and G. D. Clifford, "Screening for cognitive impairment during sleep studies: The george b. moody PhysioNet challenge 2026," https://moody-challenge.physionet.org/2026/, 2026, accessed: 2026-08-17.

[9] H. Sun, W. Ganglberger, S. Nasiri, A. Gupta, M. Ghanta, V. Moura Junior, S. Cash, K. Stone, Z. Zhang, G. Ganjoo, T. E. Nassi, R. Wei, E. Meulenbrugge, R. Au, G. Clifford, L. M. Trotti, D. Hwang, E. Mignot, U. Katwa, and M. B. Westover, "The human sleep project (version 2.0)," Brain Data Science Platform, 2023.

[10] S. Nasiri, W. Ganglberger, T. Nassi *et al.*, "CAISR: achieving human-level performance in automated sleep analysis across all clinical sleep metrics," *Sleep*, vol. 48, no. 8, p. zsaf134, 2025.

[11] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalography and Clinical Neurophysiology*, vol. 29, no. 3, pp. 306–310, 1970.

[12] H. Sun, L. Paixao, J. T. Oliva *et al.*, "Brain age from the electroencephalogram of sleep," *Neurobiology of Aging*, vol. 74, pp. 112–120, 2019.

[13] K. Lacourse, J. Delfrate, J. Beaudry, P. Peppard, and S. C. Warby, "A sleep spindle detection algorithm that emulates human expert spindle scoring," *Journal of Neuroscience Methods*, vol. 316, pp. 3–11, 2019.

[14] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[15] A. Azarbarzin, S. A. Sands, K. L. Stone, L. Taranto-Montemurro, L. Messineo, P. I. Terrill, S. Ancoli-Israel, K. Ensrud, S. Purcell, D. P. White *et al.*, "The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the osteoporotic fractures in men study and the sleep heart health study," *European heart journal*, vol. 40, no. 14, pp. 1149–1157, 2019.

[16] A.-M. G. de Lange and J. H. Cole, "Commentary: Correction procedures in brain-age prediction," *NeuroImage: Clinical*, vol. 26, p. 102229, 2020.

[17] A. Guillot and V. Thorey, "RobustSleepNet: Transfer learning for automated sleep staging at scale," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1441–1451, 2021.

[18] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1273–1283, 1993.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

[20] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021.

[21] H. Ismail Fawaz, B. Lucas, G. Forestier *et al.*, "InceptionTime: Finding AlexNet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1936–1962, 2020.

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions." Curran Associates, Inc., 2017. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf