Late Fusion of Time-Domain and Time-Frequency Sleep EEG for Future Cognitive Impairment Prediction

Nahyun Lee[1], TaeHwan Kim[1], TaeYoung Kang[1], DongJae Lee[2], Donggun Kim[3], Jeongbin You[3], Younghyuk Kim[3], Wuming Gong[4], Seungsang Oh[3], Il-Youp Kwak[1]

[1] Chung-Ang University, Seoul, South Korea
[2] Korea University Guro Hospital, Seoul, South Korea
[3] Korea University, Seoul, South Korea
[4] University of Minnesota, Lillehei Heart Institute, Minneapolis, MN, United States

Abstract

Clinical polysomnography may contain markers of cognitive decline years before diagnosis, while heterogeneous montages and acquisition conditions complicate cross-site prediction. Team CAU_KU developed a compact late-fusion system for predicting cognitive impairment one to six years after polysomnography. Using 6,600 multicenter recordings, our pipeline standardizes six EEG derivations from variable PSG configurations and samples quality-controlled windows across the night. Time-domain and time-frequency branches are trained separately, calibrated at the subject level, and fused by probability averaging. The final system achieves an age-conditioned AUROC of 0.746 on the official Challenge validation set. These results support the value of complementary sleep EEG representations within routine polysomnography for cross-site prediction of future cognitive impairment.

1. Introduction

In neurodegenerative disorders such as Alzheimer's disease, pathological changes can begin years before cognitive impairment becomes clinically apparent [1]. Earlier identification may support timely assessment, monitoring, and intervention. Sleep provides a promising setting for screening through neural oscillations, memory consolidation, and changes in autonomic and respiratory physiology [2–4]. Polysomnography records these processes using synchronized neural, cardiac, muscular, and respiratory signals [4]. Routine sleep studies may therefore contain markers associated with future dementia risk [5, 6].

The George B. Moody PhysioNet Challenge 2026 formulates this problem as a multi-site prediction task [7]. Given a clinical polysomnogram and basic demographic information, an algorithm predicts whether a patient receives a qualifying diagnosis of mild cognitive impairment, Alzheimer disease, or dementia within one to six years. The primary Challenge metric is age-conditioned AUROC, which emphasizes discrimination between patients of similar ages.

We develop a compact late ensemble using six standardized EEG derivations from clinical polysomnography. A site-robust preprocessing pipeline maps heterogeneous channel configurations to a canonical representation and samples quality-controlled windows throughout the night. Time-domain and time-frequency branches encode raw waveforms and normalized log-power spectrograms. Their subject-level outputs are calibrated and combined through probability averaging.

2. Related Work

Sleep electroencephalography and cognitive impairment. Sleep macrostructure and EEG microstructure changes are associated with cognitive decline, including reduced slow-wave sleep, altered delta and theta activity, reduced sigma power, REM slowing, and altered NREM spectral power and spindle properties [8–10]. Feature-based models using sleep architecture, spectral power, coherence, spindles, and slow oscillations report AUROCs of 0.73 for mild cognitive impairment and 0.78 for dementia [8], motivating whole-night representations that preserve spectral slowing and transient oscillatory activity.

Deep learning for cognitive screening. Deep learning studies combine EEG features and time-frequency representations with recurrent and convolutional networks to detect mild cognitive impairment and Alzheimer's disease [11–13], though most target current impairment in small, diagnosis-specific cohorts, leaving prospective prediction and cross-site generalization less explored. Larger-scale efforts include an end-to-end model trained on approximately 36,000 polysomnograms and SleepFM, trained on more than 585,000 hours of PSG for future dis-

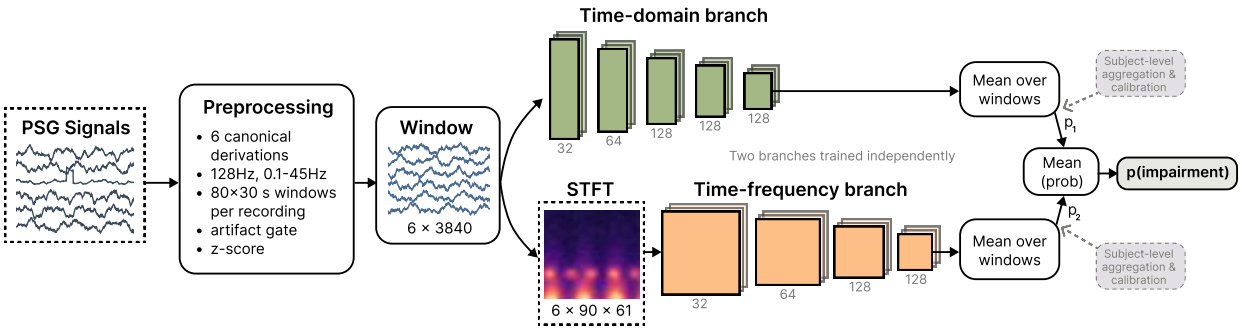

Figure 1. **Proposed late-fusion framework for future cognitive impairment prediction.** Independently trained time-domain and time-frequency encoders aggregate whole-night EEG windows into calibrated probabilities for late fusion.

ease prediction [6, 14]. The Human Sleep Project and PhysioNet Challenge extend this setting to heterogeneous multicenter PSG and future cognitive impairment at unseen institutions [7, 15].

3. Methods

3.1. EEG Preprocessing

European Data Format channel names are standardized using an alias table. We construct six canonical inputs, `F3-M2`, `F4-M1`, `C3-M2`, `C4-M1`, `O1-M2`, and `O2-M1`, each populated with the first available signal from a recorded or reconstructed bipolar derivation, the corresponding AVG-referenced channel, or the unreferenced electrode, in that order. Unavailable inputs are zero-filled and marked invalid.

Signals are resampled to 128 Hz and bandpass filtered between 0.1 and 45 Hz with a fourth-order zero-phase filter. We estimate a robust scale per channel from the full recording and sample candidate 30 s windows uniformly across the night. Within each window, a channel is flagged invalid if its peak exceeds 25 times this scale or its standard deviation falls below 5% of it, and windows with fewer than four valid channels are discarded when enough higher-quality windows remain. Each valid channel is then standardized within the window and clipped to $[-20, 20]$, giving an input tensor of shape 6×3840. This procedure limits sensitivity to signal scale, referencing, and isolated artifacts while preserving temporal and spectral patterns.

3.2. Dual-Branch EEG Encoders

Figure 1 shows the two branches described below.

Time-domain encoder. The time-domain encoder is a five-block one-dimensional convolutional network. The blocks contain 32, 64, 128, 128, and 128 output channels with kernel sizes of 15, 9, 7, 5, and 3, respectively. Each convolution uses a stride of 2, followed by group normalization and a ReLU activation. Adaptive average pooling produces a 128-dimensional window representation, and a linear layer produces one logit per window.

Time-frequency encoder. A centered short-time Fourier transform is applied independently to each channel using a 2 s Hann window, a 0.5 s hop, and an FFT size of 256. At a sampling frequency of 128 Hz, this configuration provides a frequency resolution of 0.5 Hz. The squared magnitude is restricted to 0.5–45 Hz and converted to log power as $\log_{10}(|X|^2 + 10^{-10})$. Reflection padding produces a 90×61 time-frequency representation for each channel.

Each channel spectrogram is standardized across its time-frequency plane and stacked as an input to a four-block two-dimensional convolutional network. The blocks contain 32, 64, 128, and 128 output feature maps with kernel sizes of 7×7, 5×5, 3×3, and 3×3, respectively. Each convolution uses a stride of 2, followed by group normalization and a ReLU activation. Global adaptive average pooling and a linear layer produce one logit per window.

3.3. Fusion and Decision Rule

The time-domain and time-frequency encoders are optimized separately and stored in individual checkpoints. For each encoder, window-level logits are averaged across the recording to obtain a subject-level logit. A branch-specific Platt scaling model is fitted using the internal validation subjects to convert the logit into a calibrated probability. The final continuous prediction is the arithmetic mean of the two calibrated probabilities.

Local outcome prevalence is estimated using development subjects whose ages are within two years of the target age. Each branch scales this prevalence using a multiplier selected on the internal validation set according to the Challenge reward. Selection is regularized toward the unscaled prevalence threshold to reduce overfitting. Each branch produces a positive decision when its calibrated probability exceeds the resulting threshold. The final binary prediction is positive only when both branches produce positive decisions.

	Site held out AC-AUROC		Official Challenge validation						
Model configuration	I0002	I0006	Reward	AC-AUROC	AW-AUROC	AUROC	AUPRC	Accuracy	F1
Time-frequency branch only	0.692	0.714	0.025	0.678	0.624	0.789	0.300	0.605	0.108
+ time-domain branch	0.731	**0.729**	-0.033	0.698	0.625	0.808	0.314	0.524	0.081
+ separate-process training and branch agreement	**0.739**	0.719	**0.118**	**0.746**	**0.701**	**0.836**	**0.352**	**0.665**	**0.134**

Table 1. **Performance of the submitted configurations on the large site-held-out cohorts and the official Challenge validation set.** AC-AUROC denotes age-conditioned AUROC and AW-AUROC denotes age-weighted AUROC.

4. Experiment

4.1. Dataset Description

The George B. Moody PhysioNet Challenge 2026 provides 6,600 de-identified clinical polysomnography recordings from three U.S. institutions, with hidden data from two additional institutions [16]. Recordings contain variable subsets of EEG, EOG, EMG, ECG, and respiratory signals. A positive outcome requires at least two CI-related diagnoses at least one week apart, with the first occurring one to six years after polysomnography. A negative outcome requires no CI-related diagnosis at any time and at least six years of clinical follow-up.

4.2. Training Configuration

Each recording contributes 80 evenly distributed 30 s windows. Training proceeds for 15 epochs using the AdamW optimizer with a fixed learning rate of 1×10^{-3}, a weight decay of 1×10^{-2}, and a batch size of 512 windows. The objective is weighted binary cross-entropy, with the positive-class weight defined as the ratio of negative to positive subjects. The gradient norm is clipped at 5.0. Within each training cohort, 15% of the subjects form an internal validation set stratified by outcome and age decade. The checkpoint with the highest validation AC-AUROC is retained for each branch.

We evaluate three submitted configurations. The first configuration contains only the time-frequency encoder. The second configuration trains the time-domain and time-frequency encoders within a single process using a common optimizer and separate branch losses. The branches share no parameters, and each branch retains its own best epoch and Platt calibration. Their calibrated probabilities are averaged, and a single age-dependent threshold is applied to the fused probability. The final configuration trains the two encoders in separate processes using individual optimizers and checkpoints. Their calibrated probabilities are still averaged for the continuous prediction. For the binary prediction, each branch applies its own age-dependent threshold, and the final output is positive only when both branch decisions are positive.

4.3. Evaluation Protocol and Metrics

The public training data are released in large and small partitions. Table 2 reports the composition of sites I0002 and I0006 in both partitions. All local experiments use the large partition. For local site-held-out evaluation, sites I0002 and I0006 are alternately used exclusively as test cohorts. Subjects from the remaining institutions in the large partition form the local training pool. Subjects from each held-out site do not contribute to model fitting, checkpoint selection, or probability calibration.

Held-out site	Large partition	Small partition
I0002	319 (52 pos, 16.3%)	54 (8 pos, 14.8%)
I0006	1,142 (112 pos, 9.8%)	192 (20 pos, 10.4%)

Table 2. **Composition of the local site-held-out test cohorts.** Each entry reports the total number of subjects followed by the number and percentage of positive subjects.

We separately evaluate the three submitted configurations on the official Challenge validation set. **Age-conditioned AUROC (AC-AUROC)** serves as the primary Challenge metric. It evaluates the ranking of positive and negative patients whose ages differ by no more than two years. We also report age-weighted AUROC, conventional AUROC, and AUPRC using continuous prediction scores. Reward, accuracy, and F1 are calculated from binary predictions. The reward accounts for outcome prevalence near each patient age and assigns greater value to correct predictions of less common outcomes. Each incorrect prediction receives a fixed penalty.

4.4. Main Result

Table 1 summarizes the three cumulative submission configurations. Adding the time-domain branch improves all continuous ranking metrics on the official Challenge validation set, including an increase in AC-AUROC from 0.678 to 0.698. However, reward decreases from 0.025 to -0.033, while accuracy and F1 also decline. This divergence shows that improved probability ranking alone does not ensure a reliable binary operating point and highlights the sensitivity of threshold-dependent metrics to calibration and cohort prevalence. The final configuration achieves the highest value for every official metric, including an AC-AUROC of 0.746 and a reward of 0.118. These results show that the final system improves both continuous ranking and its binary operating point.

The local results show a more variable pattern. Relative to the preceding configuration, the final configuration increases AC-AUROC from 0.731 to 0.739 on I0002 but de-

creases it from 0.729 to 0.719 on I0006. This site-specific pattern underscores the value of site-held-out evaluation during development.

5. Discussion

The time-domain branch improves every continuous ranking metric, consistent with complementary temporal and spectral information for cognitive-risk ranking; its site-specific variability suggests that the benefit of separate optimization may depend on cohort composition and training stochasticity.

The final configuration also narrows the gap between conventional AUROC and AC-AUROC, consistent with better discrimination among similarly aged patients and less reliance on broad age-related risk differences. Age is excluded from both neural networks and used only for validation stratification, AC-AUROC-based checkpoint selection, and binary thresholding. This configuration differs from the preceding one in two respects, branch optimization and binary decision formation. Because the agreement rule does not affect continuous predictions, its continuous-metric gains can be attributed to separately trained branch scores. Improvements in reward, accuracy, and F1 may also reflect branch-specific thresholds and the conservative agreement rule, so the submitted-system comparison cannot attribute binary-metric gains solely to separate training. Requiring both branches to predict positive may suppress uncertain positives and contribute to the higher reward. The relatively low F1 is consistent with this conservative operating point and possibly reduced sensitivity.

6. Conclusion

We present a compact EEG late ensemble for predicting future cognitive impairment from clinical polysomnography. The method combines canonical channel construction, quality controlled whole-night sampling, complementary time-domain and time-frequency representations, branch-specific calibration, and subject-level probability fusion. The final system achieves an AC-AUROC of 0.746 on the official Challenge validation set. These results support the feasibility of EEG-based cognitive risk prediction across heterogeneous clinical recording environments. Further evaluation with repeated training runs and additional external cohorts is required to establish the stability and generalizability of the observed improvements.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00341075, RS-2026-25477127).

References

[1] Sperling RA, et al. Toward defining the preclinical stages of alzheimer's disease. Alzheimers Dementia 2011;7(3):280–292.

[2] Rasch B, Born J. About sleep's role in memory. Physiological Reviews 2013;93(2):681–766.

[3] de Zambotti M, Trinder J, Silvani A, Colrain IM, Baker FC. Dynamic coupling between the central and autonomic nervous systems during sleep. Neuroscience Biobehavioral Reviews 2018;90:84–103.

[4] Rundo JV, Downey Ralph I. Polysomnography. Handbook of Clinical Neurology 2019;160:381–392.

[5] Pase MP, et al. Sleep architecture and the risk of incident dementia in the community. Neurology 2017;89(12):1244–1250.

[6] Thapa R, et al. A multimodal sleep foundation model for disease prediction. Nature Medicine 2026;32:752–762.

[7] George B. Moody PhysioNet Challenge. Screening for cognitive impairment during sleep studies. the george b. moody physionet challenge 2026. PhysioNet, 2026.

[8] Ye EM, et al. Dementia detection from brain activity during sleep. Sleep 2023;46:zsac286.

[9] D'Atri A, et al. Electroencephalographic alterations during wake and sleep in mild cognitive impairment and alzheimer's disease. iScience 2021;24:102386.

[10] Taillard J, et al. Non-rem sleep characteristics predict early cognitive impairment in an aging population. Frontiers in Neurology 2019;10:197.

[11] Geng D, Wang C, Fu Z, Zhang Y, Yang K, An H. Sleep EEG-based approach to detect mild cognitive impairment. Frontiers in Aging Neuroscience 2022;14:865558.

[12] Parajuli M, Amara AW, Shaban M. Deep-learning detection of mild cognitive impairment from sleep electroencephalography for patients with parkinson's disease. PLOS ONE 2023;18(8):e0286506.

[13] Gallego-Viñarás L, Mira-Tomás JM, Gaeta AM, Pinol-Ripoll G, Barbé F, Olmos PM, Muñoz-Barrutia A. Alzheimer's disease detection in EEG sleep signals. IEEE Journal of Biomedical and Health Informatics 2025;29(2):948–959.

[14] Ganglberger W, Sun H, Turley N, Tripathi A, et al. Brain health from sleep EEG: A multicohort, deep learning biomarker for cognition, disease, and mortality. NEJM AI 2026;3(3).

[15] Li Q, Wen S, Sun H, Ganglberger W, et al. The human sleep project: A multi-center clinical polysomnography dataset across the human lifespan. Sleep aug 2026;zsag215.

[16] Sun H, Ganglberger W, Nasiri S, et al. The human sleep project, 2023.

Address for correspondence:

Il-Youp Kwak
ikwak2@cau.ac.kr
DongJae Lee
konnect714@gmail.com