Fusion of Age-Residualized Engineered Sleep Features with Raw EEG for Predicting Future Cognitive Impairment

CHA KABMUN, CHOI JOUNG BAE, CHO HYO YEON

NLICA Inc, Yongin, Republic of Korea

Abstract

Overnight polysomnography may contain signatures of cognitive impairment diagnosed years later. Team NLICA developed a hybrid system to predict a diagnosis recorded 1–6 years after polysomnography, using 6,600 studies from three institutions. An engineered branch combined demographics, stage-specific EEG features, and automated-annotation descriptors. After site-wise location–scale harmonization, age was removed as a direct predictor and negative-class linear age trends were subtracted from the remaining 272 numeric features. A selected subset was modeled by a 30-member tree ensemble. In parallel, a compact one-dimensional convolutional network scored artifact-screened N2, N3, and REM epochs from four EEG derivations. Equal-weight late fusion combined the branches. Across three leave-one-site-out folds, the mean and worst-site age-conditioned AUROCs were 0.694 and 0.654. On the official hidden validation set, the age-conditioned AUROC was 0.736 and the overall AUROC 0.835. These results support cross-source predictive complementarity after age adjustment.

1. Introduction

Sleep supports cognition, and several of its electrophysiological signatures degrade with age and in neurodegenerative disease. Slow-wave sleep supports memory consolidation, and its longitudinal decline has been associated with incident dementia [1,2]. The temporal coupling of sleep spindles to slow oscillations has been proposed as a marker of the integrity of the circuits supporting overnight memory retention [3]. Because polysomnography (PSG) is already acquired at scale for the assessment of sleep-disordered breathing, it offers an accessible substrate for detecting neurodegenerative risk before a diagnosis is recorded.

The George B. Moody PhysioNet Challenge 2026 [4] posed this question directly: from an overnight PSG and basic demographics, predict whether a cognitive impairment diagnosis will be recorded between one and six years later. The task raised two central challenges.

The first is age confounding. Both the electrophysiological changes of interest and the outcome rise with age, so a model can score well on conventional AUROC by learning age alone. The Challenge ranks entries on an age-conditioned AUROC that restricts comparisons to positive–negative pairs differing by at most two years, which substantially limits the value of discrimination driven by chronological age. Reducing direct age dependence was therefore a design priority aligned with the scoring metric.

The second is cross-source generalization. Training data come from three institutions; the hidden validation and test sets come from two further institutions. Acquisition hardware, montages, filter settings, and case mix all differ. We therefore combined two branches with complementary weaknesses: a physiologically motivated feature set, harmonized across sites and adjusted for age, and a compact raw-EEG convolutional network that learns morphology the feature set does not name.

2. Methods

2.1. Data

We used the large Challenge training set: 6,600 overnight PSG recordings from three institutions, S0001 (n = 5,139), I0006 (n = 1,142), and I0002 (n = 319). Positive cases had at least two qualifying diagnoses of mild cognitive impairment, Alzheimer's disease, or dementia separated by at least one week, with the first occurring one to six years after the sleep study. Negative cases had no qualifying diagnosis at any time and at least six years of subsequent clinical encounters. Positives numbered 498 (7.5%), distributed 334, 112, and 52 across the three sites.

Each recording is accompanied by demographics and by automated annotations from the CAISR pipeline [5], which supplies 30-s sleep stages with class posteriors together with arousal, respiratory, and limb movement events. Expert human annotations exist in the training data but not in the hidden sets, so we excluded them entirely, keeping the training and inference feature spaces

identical.

Unipolar channels were re-referenced to form six bipolar EEG derivations — F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, and O2-M1 — wherever the constituent channels shared a sampling rate. The engineered branch used all six; the convolutional branch used the central and occipital subset.

2.2. Engineered feature branch

EEG signals were resampled to 128 Hz, notch filtered, bandpass filtered at 0.3–35 Hz, amplitude clipped, and standardized by median and median absolute deviation. A 30-s epoch was rejected if its peak-to-peak range exceeded five times the recording's median epoch range or its standard deviation fell below 10^{-4}.

Each recording was then reduced to 281 features. Demographics contributed 10: age, one-hot sex and race, and body mass index. EEG-derived features contributed 186, all computed within sleep stages rather than over the whole night, because the physiological meaning of a rhythm depends on state: stage-specific spectra (6 derivations × 20), spindle density, mean amplitude, sigma-band power and spindle occupancy (6 × 4), alpha peak frequency (6 × 2), slow oscillation descriptors (6 × 4), and slow-oscillation–spindle coupling (6 × 1). Coupling was quantified within N2 as a mean vector length modulation index between the 0.5–1.25 Hz phase and the 11–16 Hz envelope, both obtained by Hilbert transform of bandpass-filtered signals. Summaries of the CAISR annotations contributed 85, comprising standard indices — apnea-hypopnea, arousal, periodic limb movement, stage percentages, sleep efficiency — and descriptors of how sleep architecture evolves across the night. Non-finite values were replaced with zero; no explicit missingness indicators were used.

After the adjustments described in Section 2.3, random-forest impurity importance retained the 200 most important features. These were standardized and passed to an ensemble of 30 models — XGBoost [6], random forest [7], and extremely randomized trees [8] at each of 10 seeds — whose predicted probabilities were averaged.

2.3. Site harmonization and age adjustment

Two transformations preceded feature selection. Both were fitted on training data only, and both were re-fitted inside every cross-validation fold.

For site-wise location–scale harmonization, we standardized each site's records by its own per-feature mean and standard deviation, then mapped them onto a common scale defined by the pooled training statistics. This is a location–scale batch adjustment in the spirit of ComBat [9] but without covariate modeling or empirical-Bayes shrinkage. For unseen sites no site statistics exist, so a fixed reference formed by averaging the available I-series training-site statistics was used, matching the submitted inference procedure. Leave-one-site-out folds treat the held-out site the same way, so the evaluation reproduces the deployment path rather than an optimistic one.

The combined age adjustment applied two operations together. The explicit age column was set to zero, removing age as a direct predictor. Then, using only negative-class training records, we estimated by ordinary least squares the linear slope β of each remaining numeric feature x on chronological age a and subtracted the fitted trend from every record, $x' = x - \beta(a - \bar{a})$. Fitting on negative-class records only is the substantive choice: it defines a reference aging trajectory among non-cases, so that the residual expresses deviation from that reference rather than age itself. Fitting on all records could partly absorb case-related variation that is correlated with age. The subtraction covered all 272 numerically encoded features from index 9 onward, leaving the one-hot sex and race indicators untouched, since a linear age adjustment of a binary indicator is not interpretable.

Ordinary least squares removes the linear component by construction; monotonic association is only attenuated. Among the negative-class records used to fit each fold's adjustment, the median absolute Spearman correlation between age and the adjusted features fell from 0.076–0.082 to 0.013–0.017, a reduction of 78.0–84.5%, and the 90th percentile fell from 0.165–0.198 to 0.093–0.118. These are in-sample diagnostics and indicate substantial but incomplete attenuation.

2.4. Raw-EEG convolutional branch

In parallel, a compact one-dimensional convolutional network operated on the signal itself. Input was four channels (C3-M2, C4-M1, O1-M2, O2-M1) of 30-s epochs at 128 Hz, standardized per channel over time, drawn from N2, N3, and REM as labeled by CAISR and capped at 20 epochs per stage per recording. Only these stages were used, to reduce movement contamination and stage ambiguity.

The backbone comprised four blocks of convolution, batch normalization, GELU activation, and dropout, with 32, 48, 64, and 96 output channels at decreasing temporal resolution, followed by global average pooling and a single linear unit — approximately 90,000 parameters. The small capacity was chosen to limit overfitting given only 498 positive patients. Training used epoch-level binary cross-entropy with the patient's label broadcast to all of that patient's epochs and a positive-class weight set to the epoch-level negative-to-positive ratio, optimized by AdamW at a learning rate of 3×10^{-4} for a fixed budget of 25 epochs. At inference, epoch probabilities were averaged across all retained N2, N3, and REM epochs for

each patient. A CNN prediction was available for 6,521 of 6,600 patients (98.8%).

2.5. Fusion, decision rule, and evaluation

Branch probabilities were combined by fixed equal-weight late fusion, p = 0.5 p_eng + 0.5 p_CNN, with the engineered-branch probability used unchanged when no CNN prediction was available. Equal weighting was applied as a simple fixed rule and was not optimized against the hidden validation set; with three training sites, a weight chosen to maximize worst-site performance is fitted to three numbers.

Binary predictions, required by the accuracy, F-measure, and reward metrics, were generated by labeling the highest-scoring 10% of the cohort positive; the rate was fixed at 10% after a training-set sweep between 5% and 15%. The hidden sets have a stated prevalence in that range, and a rank-based rule avoids reliance on absolute probability calibration at an unseen site. The rule affects only the binary metrics; the ranking used by both AUROC metrics is unchanged.

We evaluated by leave-one-site-out cross-validation, holding out one institution at a time and re-estimating every fitted component — harmonization statistics, age-adjustment coefficients, feature selection, scaling, and both branches — inside each fold. Scores were computed with the official implementation at a gap of two years. We summarize by the unweighted mean across held-out sites and by the minimum, treating the minimum as the primary internal criterion because it is the closest available proxy for an unseen institution.

3. Results

Table 1. Leave-one-site-out age-conditioned AUROC (gap = 2 years) by branch and after fusion.

Site	Engineered	CNN	Fused
I0002	0.719	0.708	0.726
I0006	0.703	0.677	0.703
S0001	0.642	0.639	0.654
Mean	0.688	0.675	0.694
Worst	0.642	0.639	0.654

CNN-only values were computed among patients with an available CNN output; engineered and fused values used the full held-out cohorts, with engineered fallback where no CNN output existed. Fusion raised the worst-site age-conditioned AUROC from 0.642 to 0.654 and the site-equal mean from 0.688 to 0.694. Fusion effects were heterogeneous across sites: the largest gain occurred at S0001, where the engineered branch was weakest, while at I0006 fusion produced no net change despite a 0.026 difference between the branches. S0001 was also the most demanding fold to fit — holding it out leaves 1,461 training patients and 1,297 negative-class records from which to estimate the age-adjustment coefficients, against 5,835 when I0002 is held out.

Table 2. Hidden-validation performance of the original system (entry 2155) and the combined age-adjusted system (entry 2361).

Metric	2155	2361	Δ
Age-cond. AUROC	0.720	0.736	+0.016
Age-wt. AUROC	0.690	0.715	+0.025
AUROC	0.836	0.835	−0.001
AUPRC	0.322	0.302	−0.020
Reward	0.146	0.172	+0.026

The combined age adjustment raised the age-conditioned AUROC from 0.720 to 0.736 and the age-weighted AUROC from 0.690 to 0.715, while overall AUROC changed by −0.001. AUPRC decreased by 0.020, indicating a trade-off in precision–recall performance. Since the Challenge ranks on the age-conditioned metric, we accepted that trade.

Hidden-validation performance exceeded leave-one-site-out performance (0.736 against a mean of 0.694). This may reflect the larger training sample available to the submitted model — each fold withholds an entire institution, and the S0001 fold trains on 1,461 patients against 6,600 for the submission — as well as differences in cohort composition, prevalence, and age distribution between the validation institution and the training sites.

4. Discussion

The combined age adjustment improved both age-aware ranking metrics on an institution absent from training while overall AUROC changed by only −0.001; the accompanying decrease in AUPRC indicates a trade-off in precision–recall performance rather than selective removal of age-only information. These results are consistent with reduced reliance on direct and linear age-related information, but they do not establish age invariance.

The adjustment is linear and applied uniformly, and both choices are limitations. Ordinary least squares removes only the linear component; our in-sample diagnostics show median monotonic association attenuated by roughly 80% but not eliminated. More importantly, the same continuous subtraction was applied to genuinely continuous features and to low-cardinality numeric features alike. For the latter, subtracting an age-dependent term can break ties and order previously indistinguishable records by age, increasing rather than decreasing rank association. Future work should use spline or generalized additive normative models [10] for continuous variables, suitable models for counts and bounded proportions, and exclude ordinal variables from continuous residualization.

Per-site evaluation showed heterogeneous fusion effects, with the largest gain at S0001, where the engineered branch was weakest. We therefore used the per-site vector and its minimum for internal model selection, noting that the Challenge itself ranks on a single hidden test institution rather than on any cross-validation statistic.

Two further limitations bear on interpretation. The convolutional branch is trained with the patient-level label broadcast to every epoch, so many epochs in a positive recording may carry little disease signal, introducing substantial label noise; multiple-instance formulations are a natural alternative. And all conclusions rest on three training institutions and one hidden validation institution.

5. Conclusion

An overnight polysomnogram carries information about a cognitive impairment diagnosis recorded one to six years later, and predictive value remained after removal of chronological age as a direct predictor and linear age adjustment of the engineered features. Combining harmonized, age-adjusted sleep features with a compact raw-EEG convolutional branch achieved an age-conditioned AUROC of 0.736 on a hidden validation set from an unseen institution. Whether this holds on a further institution is the question the Challenge test set will answer.

Conflict of Interest

All authors are co-founders of NLICA Inc; CHA KABMUN serves as its chief executive. The Challenge data were provided by the organizers, who had no role in the design or reporting of this work.

References

[1] Diekelmann S, Born J. The memory function of sleep. Nat Rev Neurosci 2010;11(2):114–126.

[2] Himali JJ, Baril AA, Cavuoto MG, Yiallourou S, Wiedner CD, Himali D, et al. Association between slow-wave sleep loss and incident dementia. JAMA Neurol 2023;80(12):1326–1333.

[3] Helfrich RF, Mander BA, Jagust WJ, Knight RT, Walker MP. Old brains come uncoupled in sleep: slow wave-spindle synchrony, brain atrophy, and forgetting. Neuron 2018;97(1):221–230.e4.

[4] George B. Moody PhysioNet Challenge. Screening for cognitive impairment during sleep studies: the George B. Moody PhysioNet Challenge 2026. 2026. Available at: https://moody-challenge.physionet.org/2026/. Accessed 13 August 2026.

[5] Nasiri S, Ganglberger W, Nassi TE, Meulenbrugge EJ, Moura Junior V, Ghanta M, et al. CAISR: achieving human-level performance in automated sleep analysis across all clinical sleep metrics. Sleep 2025;48(8):zsaf134.

[6] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proc KDD 2016;785–794.

[7] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[8] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42.

[9] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8(1):118–127.

[10] Marquand AF, Kia SM, Zabihi M, Wolfers T, Buitelaar JK, Beckmann CF. Conceptualizing mental disorders as deviations from normative functioning. Mol Psychiatry 2019;24(10):1415–1424.

Address for correspondence:

CHA KABMUN
NLICA Inc
Rm 407, Global Industry-Academic Cooperation Bldg
Dankook University Jukjeon Campus
152 Jukjeon-ro, Suji-gu, Yongin-si 16890
Gyeonggi-do
Republic of Korea
https://www.nlica.com
indie486@naver.com