Integrating Sleep Foundation Model Representations and Digital Biomarkers for Age-Aware Cognitive Risk Prediction

Leyu Zhou[1,2], Donglin Xie[2,3], Guangkun Nie[2,4], Jinshuai Gu[2,5], Qingshuo Zhao[5], Shijie Chen[6], Deyun Zhang[6], Yuxi Zhou[5], Shenda Hong[2,3,7]

[1]School of Artificial Intelligence, University of Science and Technology Beijing, Beijing, China
[2]National Institute of Health Data Science, Peking University, Beijing, China
[3]Institute of Medical Technology, Peking University Health Science Center, Beijing, China
[4]School of Intelligence Science and Technology, Peking University, Beijing, China
[5]Department of Computer Science, Tianjin University of Technology, Tianjin, China
[6]HeartVoice Medical Technology, Hefei, China
[7]Institute for Artificial Intelligence, Peking University, Beijing, China

Abstract

Predicting future cognitive impairment from polysomnography (PSG) is challenging because sleep physiology varies strongly with age and across clinical centers. We integrated pretrained SleepFM representations with compact and extended digital sleep biomarkers, demographic information, and age-matched pairwise ranking in a unified prediction framework. The submitted system achieved an official validation age-conditioned AUROC of 0.626. Among candidate sleep biomarkers, transition instability showed consistent associations with cognitive impairment across all three development centers and a positive adjusted association after accounting for age, sex, BMI, and site, whereas the respiratory effort-related arousal (RERA) index showed marked center dependence. These results support combining foundation-model representations with robust digital biomarkers for age-aware multicenter cognitive-risk prediction.

1. Introduction

Sleep abnormalities have been increasingly associated with cognitive decline and neurodegenerative disease. Changes in sleep architecture, arousals, respiratory physiology, and electroencephalographic activity may precede clinical cognitive impairment, making polysomnography (PSG) a promising source of prognostic information [1–3]. However, multicenter PSG prediction remains challenging because sleep physiology and cognitive risk vary strongly with age, while recording configurations, annotations, and patient populations vary across centers.

This task therefore requires addressing two related problems. First, conventional classification may exploit age-associated differences rather than distinguish participants with different cognitive outcomes at similar ages. Age-conditioned AUROC evaluates positive–negative pairs only when their ages differ by no more than two years, motivating learning that reflects age-matched discrimination. Second, sleep biomarkers may not transfer consistently across centers because their measurements depend on acquisition and scoring procedures. Robust prediction therefore requires age-aware learning and representations informative across heterogeneous environments.

PSG contains both high-dimensional physiological patterns and interpretable sleep characteristics. Pretrained models such as SleepFM capture latent information from raw recordings [4], while sleep architecture summarizes clinically meaningful organization associated with cognitive function [5]. Transition-related uncertainty further describes instability between sleep states. We therefore developed a framework integrating pretrained PSG representations, digital sleep biomarkers, and demographic variables through a gradient-boosted meta-classifier. An age-matched pairwise ranking model was incorporated to better align prediction with age-conditioned discrimination, while candidate biomarkers were compared across centers to assess transportability.

2. Methods

The data were derived from the Human Sleep Project [6], with automated sleep-stage and event annotations provided by CAISR [7]. Our framework integrates multi-source sleep representations with a pointwise gradient-boosted meta-classifier and an age-matched pairwise ranking model, as illustrated in Fig. 1.

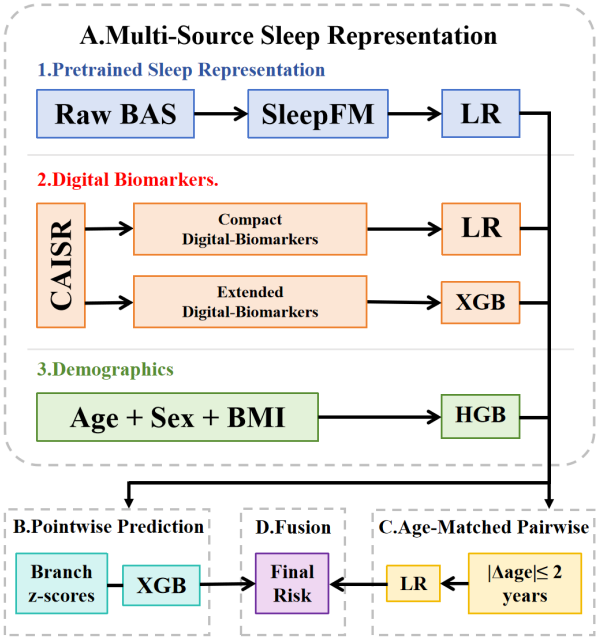

Figure 1. Overview of the prediction framework. (A) Four branches extract pretrained SleepFM representations, compact and extended digital sleep biomarkers derived from CAISR annotations, and demographic information. (B) Pointwise XGBoost combines branch z-scores. (C) Age-matched pairwise logistic regression uses branch-score differences from pairs with age difference ≤ 2 years. (D) The pointwise and ranking scores are fused into the final risk score. BAS, brain-activity signals; LR, logistic regression; XGB, XGBoost; HGB, histogram-based gradient boosting.

2.1. Multi-Source Sleep Representation and Pointwise Fusion

As shown in Fig. 1A, four complementary branches were constructed from raw PSG signals, automated sleep-stage annotations, and demographic information. The two annotation-based branches formed compact and extended digital sleep-biomarker representations.

Pretrained sleep representation. Brain-activity signals were harmonized toward six canonical derivation slots: F3-M2, C3-M2, O1-M2, F4-M1, C4-M1, and O2-M1. Mastoid-referenced derivations were preferred; auricular-referenced and, when necessary, average-referenced channels were used as surrogate fallbacks to improve coverage across heterogeneous montages. At least three usable channels were required. Signals were resampled to 128 Hz and processed with the pretrained SleepFM BAS encoder. Valid 128-dimensional embeddings were mean-pooled to a recording-level representation and mapped to a risk score using standardized logistic regression ($C = 0.5$).

Compact digital-biomarker representation. The digital biomarkers used in the prediction branches were derived from CAISR-generated automated sleep-stage outputs rather than manually scored annotations. The compact representation included N2 fraction and the 90th percentile of stage-probability entropy. For epoch t,

$$H_t = -\sum_{k=1}^{5} p_{t,k} \ln p_{t,k}, \tag{1}$$

where $p_{t,k}$ is the normalized probability of stage k. The two features were modeled using a robust scaler and logistic regression ($C = 1.0$).

Extended digital-biomarker representation. The extended representation combined broader sleep-architecture biomarkers with transition-related measures of sleep-state instability. It included N1 and REM duration, N2 fraction, a sleep-efficiency proxy defined as the proportion of valid staged epochs not labeled as Wake, mean stage-probability entropy, and the mean and maximum entropy computed separately over the first and second halves of the recording.

Because CAISR stages are defined at 30-s resolution, for a temporal neighborhood of w seconds we set $r_w = w/30$ epochs and defined

$$T_t^{(w)} = I\left(|\{y_\tau : |\tau - t| \leq r_w\}| > 1\right). \tag{2}$$

Thus, $w = 30$ and 60 s correspond to one and two neighboring epochs on each side, respectively. Two transition measures were retained: the fraction of 30-s transition-related epochs with $H_t > 1.0$ nat,

$$F_{\mathrm{HE30}} = \frac{\sum_t T_t^{(30)} I(H_t > 1.0)}{\sum_t T_t^{(30)}}, \tag{3}$$

and the maximum entropy among 60-s transition-related epochs, $H_{\mathrm{T60,max}}$. Features that could not be computed were treated as missing. The complete feature set was modeled by a shallow XGBoost classifier (maximum depth 2, learning rate 0.05, 100 trees).[8]

Demographic representation. Age, sex, and body mass index (BMI) were modeled using histogram-based gradient boosting (maximum depth 2, learning rate 0.05, 200 iterations).

Each branch produced a raw score and a standardized z-score. The four z-scores and four corresponding missingness indicators were combined using a shallow XGBoost meta-classifier (maximum depth 2, learning rate 0.03, 100 trees). Missing standardized scores were set to zero. All feature-scaling and score-standardization statistics were estimated from the training data and applied unchanged at inference.

To reduce imbalance across centers, age groups, and outcome classes, participants were grouped by site, outcome label, and four age intervals: younger than 55, 55–64, 65–74, and 75 years or older. The sample weight was defined as

$$w_i \propto \frac{1}{\sqrt{N_{g(i)}}}, \tag{4}$$

where $N_{g(i)}$ is the number of training participants in the site–age-group–class stratum containing participant i.

For the submitted system, the branch models were refitted on the full training cohort and the meta-classifier was subsequently trained on their branch scores. Its standardized output is denoted z_{point}.

2.2. Age-Matched Pairwise Ranking and Score Fusion

The pointwise model does not explicitly optimize discrimination between participants of similar ages. We therefore constructed an age-matched pairwise ranking model using the four raw branch scores,

$$\mathbf{s} = [\mathbf{s}_{\mathrm{SleepFM}}, \mathbf{s}_{\mathrm{compact}}, \mathbf{s}_{\mathrm{extended}}, \mathbf{s}_{\mathrm{demo}}]. \tag{5}$$

A cognitively impaired participant i and control participant j formed an eligible pair when

$$|a_i - a_j| \leq 2, \tag{6}$$

where a denotes age in years. Pairwise logistic regression was trained on

$$\Delta\mathbf{s_{ij}} = \mathbf{s_i^+} - \mathbf{s_j^-}, \tag{7}$$

with the reverse difference assigned to the opposite class. Eligible pairs could span training centers and were randomly subsampled to at most 50,000 pairs.

At inference, the learned linear function was applied directly to each participant's four raw branch scores and standardized to obtain z_{rank}. The final score was

$$z_{\mathrm{final}} = 0.5 z_{\mathrm{point}} + 0.5 z_{\mathrm{rank}}. \tag{8}$$

Global permutation importance was evaluated under three-site LOSO by permuting each interpretable feature within the held-out site and recomputing the affected branches and final Age-AUROC. SleepFM was treated as one 128-dimensional feature group, and N2 fraction was permuted jointly in the compact and extended branches.

2.3. Candidate Sleep Biomarkers

To examine cross-site transportability, we evaluated conventional physiological biomarkers together with algorithm-derived digital sleep biomarkers. Conventional biomarkers included RERA and arousal indices, defined as event counts per hour of sleep; REM fraction, defined as the proportion of valid staged epochs in REM; RMSSD, defined as the root mean square of successive normal-to-normal interval differences; and T90, defined as the proportion of valid overnight time with $\mathrm{SpO}_2 < 90\%$.

Among the algorithm-derived digital biomarkers, transition instability was represented by F_{HE30} defined above. All candidate biomarkers were compared under the same three-site leave-one-site-out protocol using logistic regression with inverse-square-root site–class weighting and Age-AUROC.

3. Results

3.1. Official Validation Performance

On the official hidden large validation set, the submitted system achieved an Age-AUROC of 0.626. An earlier submission variant without age-bin reweighting and the age-matched pairwise ranking component achieved an Age-AUROC of 0.595, corresponding to an absolute increase of 0.031 in Age-AUROC.

3.2. Cross-Site Transportability of Sleep Biomarkers

Candidate biomarkers showed substantial differences in cross-site transportability under the unified LOSO evaluation (Table 1). Transition instability showed the strongest overall robustness: its mean Age-AUROC was 0.683 and its worst-site Age-AUROC was 0.636. Its site-specific Age-AUROC ranged from 0.636 to 0.756, indicating relatively consistent discrimination across centers.

In contrast, event-based biomarkers were more center dependent. RERA achieved Age-AUROCs of 0.719 on I0006 and 0.688 on I0002, but only 0.482 on S0001. Arousal index showed a similar pattern, whereas REM fraction, RMSSD, and T90 provided weaker or less consistent cross-site discrimination.

Table 1. Cross-site age-conditioned AUROC (Age-AUROC) of candidate sleep biomarkers under a unified leave-one-site-out (LOSO) protocol.

Biomarker	S0001	I0002	I0006	Mean
Transition instability	0.636	0.756	0.657	0.683
RERA index	0.482	0.688	0.719	0.629
Arousal index	0.554	0.636	0.701	0.631
REM fraction	0.558	0.609	0.595	0.587
HRV (RMSSD)	0.517	0.497	0.532	0.515
Hypoxia (T90)	0.481	0.500	0.462	0.481

3.3. Transition Instability and Cognitive Impairment

We further examined F_{HE30}, an algorithm-derived digital biomarker characterizing transition-related sleep-state instability. The median 30-s high-entropy transition fraction was higher in participants who developed cognitive impairment in all three centers: 0.300 versus 0.227 in S0001, 0.325 versus 0.230 in I0002, and 0.386 versus 0.298 in I0006. The corresponding Cliff's δ values were 0.307, 0.470, and 0.338.

In multivariable logistic regression, the adjusted coefficient for transition instability remained positive after accounting for age, sex, BMI, and site (coefficient $= 2.65$ for the transition fraction on its original 0–1 scale). Moreover, among 20,390 positive–negative pairs matched within two years of age, the cognitively impaired participant had higher transition instability in 63.2% of pairs. Together, these analyses indicate that the association between transition instability and cognitive impairment was not attributable solely to chronological age.

3.4. Global Feature Importance

Permutation analysis ranked SleepFM highest in the final model (mean ΔAge-AUROC $= 0.1134$). Among interpretable features, F_{HE30} ranked highest (0.0234), followed by age (0.0190), BMI (0.0181), entropy P90 (0.0083), and second-half mean entropy (0.0068). Negative importance for lower-ranked features may reflect redundancy or interactions among correlated inputs rather than a lack of predictive information.

Table 2. Global feature importance ranked by mean ΔAge-AUROC (ΔAUC). H1/H2, first/second half; SE, sleep efficiency.

Feature	ΔAUC	Feature	ΔAUC
SleepFM	**0.1134**	Stage Ent.	0.0035
F_{HE30}	**0.0234**	H1 Ent. mean	0.0021
Age	0.0190	N1 dur.	0.0013
BMI	0.0181	Sex	0.0003
Ent. P90	0.0083	H2 Ent. max	0.0002
H2 Ent. mean	0.0068	SE	0.0000
N2 frac.	0.0063	H1 Ent. max	-0.0010
REM dur.	0.0059	$H_{\mathrm{T60,max}}$	-0.0023

4. Conclusion

We developed a multicenter PSG-based cognitive-risk prediction framework combining pretrained SleepFM representations, compact and extended digital sleep biomarkers, demographic information, and age-matched pairwise ranking. The system achieved an official validation Age-AUROC of 0.626, compared with 0.595 for an earlier submission variant without the age-aware components. Among sleep biomarkers, transition instability showed the most consistent cross-site performance and a positive age-adjusted association with cognitive impairment, whereas several event-based biomarkers showed center dependence. These results support the importance of age-aligned learning and biomarker transportability in multicenter PSG-based cognitive-risk prediction.

Acknowledgments

This work was supported by the Beijing Nova Program under Grant 202604841257.

References

[1] Bubu OM, Brannick M, Mortimer J, Umasabor-Bubu O, Sebastião YV, Wen Y, Schwartz S, Borenstein AR, Wu Y, Morgan D, Anderson WM. Sleep, cognitive impairment, and Alzheimer's disease: a systematic review and meta-analysis. Sleep 2017;40(1):zsw032.

[2] Wennberg AMV, Wu MN, Rosenberg PB, Spira AP. Sleep disturbance, cognitive decline, and dementia: A review. Seminars in Neurology 2017;37(4):395–406.

[3] Brink-Kjaer A, et al. Age estimation from sleep studies using deep learning predicts life expectancy. npj Digital Medicine 2022;5:103.

[4] Thapa R, Kjaer MR, He B, Covert I, Moore IV H, Hanif U, Ganjoo G, Westover MB, Jennum P, Brink-Kjaer A, Mignot E, Zou J. A multimodal sleep foundation model for disease prediction. Nature Medicine 2026;32:752–762.

[5] Blackwell T, Yaffe K, Ancoli-Israel S, Redline S, Ensrud KE, Stefanick ML, Laffan A, Stone KL. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. Journal of the American Geriatrics Society 2011;59(12):2217–2225.

[6] Sun H, Ganglberger W, Nasiri S, Gupta A, Ghanta M, Moura Junior V, et al. The human sleep project (version 2.0), 2023.

[7] Nasiri S, Ganglberger W, Nassi T, 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.

[8] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016; 785–794.

Address for correspondence:

Shenda Hong, National Institute of Health Data Science, Peking University, 38 Xueyuan Road, Haidian District, Beijing 100191, China; E-mail: hongshenda@pku.edu.cn