Multimodal Sleep Signal Phenotyping for Predicting Future Cognitive Impairment

Sejuti Dhali[1], Tasmeya Siddique Mayesha[1]
[1]*Department of Biomedical Engineering, Khulna University of Engineering & Technology, Khulna-9203, Bangladesh*

Abstract

Overnight polysomnography (PSG) may provide physiological evidence of future cognitive impairment; however, multicenter heterogeneity and low outcome prevalence complicate model development. Following official-phase updates to the Challenge data and scoring framework, we developed a multimodal late-fusion system using 1,103 public-training PSG records from three acquisition sites, including 84 positive outcomes (7.6%). Three random-forest branches modelled demographic features, engineered PSG-physiology features and CAISR-derived sleep phenotypes. Their site-held-out out-of-fold probabilities were combined using a logistic-regression stacker. All learned preprocessing and modelling operations - including imputation, missingness processing, dimensionality reduction, physiology site-stability filtering, branch training and stacking - were restricted to the applicable training partition under nested leave-one-site-out evaluation. Late fusion achieved pooled AUROC 0.670, AUPRC 0.162 and Brier score 0.070. The demographic baseline produced the highest AUROC (0.687); however, late fusion increased AUPRC from 0.132 to 0.162, representing a 22.7% relative improvement and approximately 2.1 times the positive-outcome prevalence. Age-conditioned AUROC was 0.473, indicating limited discrimination beyond age. These values are public-training cross-validation estimates; no official validation score, test score, or ranking was available. The finalized framework therefore improved rare-case prioritization but requires hidden-data or external multicenter evaluation before clinical interpretation.

1. Introduction

Overnight polysomnography (PSG) captures multidimensional physiological signals during sleep, offering a window into cognitive-risk assessment, since neurophysiological changes may precede clinical impairment. Sleep EEG-derived measures, including brain-age indices [1] and multivariate ambulatory EEG features [2], have been linked to incident dementia and future cognitive decline, establishing sleep EEG as a candidate prognostic source. Evidence from conventional sleep architecture is less consistent: one meta-analysis reported structural differences in mild cognitive impairment [3], while a larger five-cohort analysis found no consistent association between sleep-stage proportions and dementia [4], suggesting macrostructural summaries alone may not capture the heterogeneity needed for risk prediction.

Sleep EEG microstructure and learned representations may capture information stage-summaries obscure. REM-sleep slowing [5], spectral-spindle-slow-wave combinations [6,7], and multicohort deep-learning representations [8] have each been linked to cognitive status, though evidence remains largely association-focused. Translation into transportable models is complicated by age-related EEG change [9], risking discrimination that reflects age rather than disease, and by site-specific differences that inflate performance under conventional cross-validation [10]. Resources such as the Human Sleep Project [11], CAISR phenotyping [12], and the PhysioNet Challenge 2026 [13] motivate leakage-safe, multimodal, beyond-demographic evaluation, with precision-recall and calibration metrics needed alongside AUROC given the outcome's rarity [14,15].

This study (by Team SomnoSense) addresses these gaps using a three-branch framework on 1,103 public-training PSG records from three sites, combining demographic, engineered physiology, and CAISR-derived phenotype features through logistic late fusion. Its contributions: a multimodal representation spanning demographic, neurophysiological, autonomic, respiratory, oximetric, and phenotype information; a nested leave-one-site-out evaluation restricting preprocessing and modelling to training sites; and a reporting framework distinguishing global discrimination, rare-positive prioritization, probability error, site-specific performance, and discrimination beyond age. Results represent leakage-safe cross-validation estimates on the public-training release, not official Challenge validation or hidden-test performance.

2. Methods

2.1 Challenge cohort and prediction target

The accepted abstract reported an interim audit of an earlier 622-record training subset. Following the official-phase updates to the Challenge data all analyses presented in this paper were repeated using the complete large public-training release. The final cohort comprised 1,103 overnight PSG records from three acquisition sites: I0002 (n = 54), I0006 (n = 192) and S0001 (n = 857). The Challenge-provided binary outcome identified 84 records with future cognitive impairment and 1,019 records without the outcome, corresponding to a prevalence of 7.6%. Each record was uniquely identified by its site, BIDS folder and session. Challenge-provided labels and demographic variables were used without re-labelling, and no external participant-level data were introduced. The Human Sleep Project and PhysioNet provide the multicenter data context [11,13].

2.2 Multimodal feature representation

Each record was represented using three complementary feature families. The demographic branch received a 16-dimensional vector derived from age, sex, body-mass index, available demographic metadata and

corresponding missingness indicators. The physiology branch began with 537 candidate variables summarizing whole-night EEG spectral properties, Hjorth activity, mobility, and complexity [8], Higuchi fractal dimension [7], ECG and heart-rate variability, respiratory dynamics, oxygen saturation and channel-quality characteristics. The CAISR branch contained 189 automated sleep-phenotype variables, including sleep-stage occupancy, sleep efficiency, wake after sleep onset, stage transitions, annotation confidence, fractal summaries and latent trajectory components. CAISR outputs were treated as auxiliary phenotypic evidence rather than clinical ground truth.

2.3 Training-only preprocessing and feature stability

All learned preprocessing operations were fitted exclusively within the applicable outer-fold training partition. Numeric missing values were replaced using medians calculated from the training sites, and missingness indicators were generated without reference to the held-out site. The physiology branch additionally applied a training-only site-stability filter. A candidate physiology feature was removed if its observed value or missingness pattern predicted training-site identity with AUROC greater than 0.65, or if its association with the outcome reversed direction across the training sites with an absolute association of at least 0.05. This procedure was intended to reduce reliance on acquisition-specific artefacts rather than to identify causal biomarkers. The CAISR feature matrix was transformed using incremental principal-component analysis fitted only on training records. Consequently, held-out-site values did not influence imputation statistics, missingness processing, physiology-filter decisions, or principal components.

2.4 Branch models and out-of-fold late fusion

The demographic, physiology and CAISR feature families were modelled using separate random-forest classifiers [3]. Each classifier contained 300 trees, used balanced-subsample class weighting, and required a minimum of three training observations per terminal leaf. These hyperparameters were fixed before outer-fold evaluation; no information from the held-out acquisition site was used for model selection or tuning. The three branch probabilities were combined through L2-regularized logistic stacked generalization [4]. For record i, the final probability was defined as,

$$\text{logit}(p_i) = \beta_0 + \beta_D p_{D,i} + \beta_P p_{P,i} + \beta_C p_{C,i} \quad (1)$$

where $\text{logit}(p) = \ln\left[\frac{p}{(1-p)}\right]$, $p_{D,i}$, $p_{P,i}$, and $p_{C,i}$ denote the demographic, physiology and CAISR branch probabilities; β_0 denotes the intercept; and β_D, β_P and β_C denote the learned stacking coefficients. The logistic-regression stacker used C = 0.5 without class weighting. Its training inputs consisted exclusively of inner site-held-out out-of-fold branch probabilities, ensuring that no record contributed a stacking input generated by a branch model fitted on that record. When CAISR features were unavailable, demographic and physiology probabilities were supplied to a separately fitted two-branch fallback stacker.

3. Results

3.1 Pooled out-of-fold performance

Nested leave-one-site-out evaluation produced one outer-fold out-of-fold probability for each of the 1,103 public-training records. Figure 2 presents the corresponding receiver-operating-characteristic and precision-recall curves. The three-branch late-fusion model achieved a pooled AUROC of 0.670, an AUPRC of 0.162 and a Brier score of 0.070. The demographic baseline achieved AUROC 0.687 and AUPRC 0.132, while the physiology and CAISR branches achieved AUROC/AUPRC values of 0.614/0.133 and 0.653/0.148, respectively.

Relative to the demographic baseline, late fusion had an AUROC difference of −0.017 and an AUPRC difference of +0.030, corresponding to a 22.7% relative increase in AUPRC. The positive-outcome prevalence was 0.076; therefore, the late-fusion AUPRC was approximately 2.1 times the prevalence reference. Age-conditioned AUROC was 0.473 for late fusion, compared with 0.495 for demographics, 0.558 for physiology and 0.573 for CAISR. Table 1 summarizes the pooled branch and fusion results. The final late-fusion model achieved a pooled Brier score of 0.070.

Model	AUROC	AUPRC	Age-conditioned AUROC
Demographics (baseline)	0.687	0.132	0.495
Physiology	0.614	0.133	0.558
CAISR	0.653	0.148	0.573
Three-branch late fusion	**0.670**	**0.162**	0.473

Table 1. Pooled public-training out-of-fold model performance.

3.2 Held-out-site performance

Performance differed across the three outer folds. When I0002 was held out, the late-fusion model achieved AUROC 0.731 and age-conditioned AUROC 0.583 across 54 records. The corresponding values were 0.560 and 0.415 for the 192 records from I0006 and 0.740 and 0.491 for the 857 records from S0001. Pooling the outer-fold predictions yielded AUROC 0.670 and age-conditioned AUROC 0.473. Table 2 reports the site-specific values.

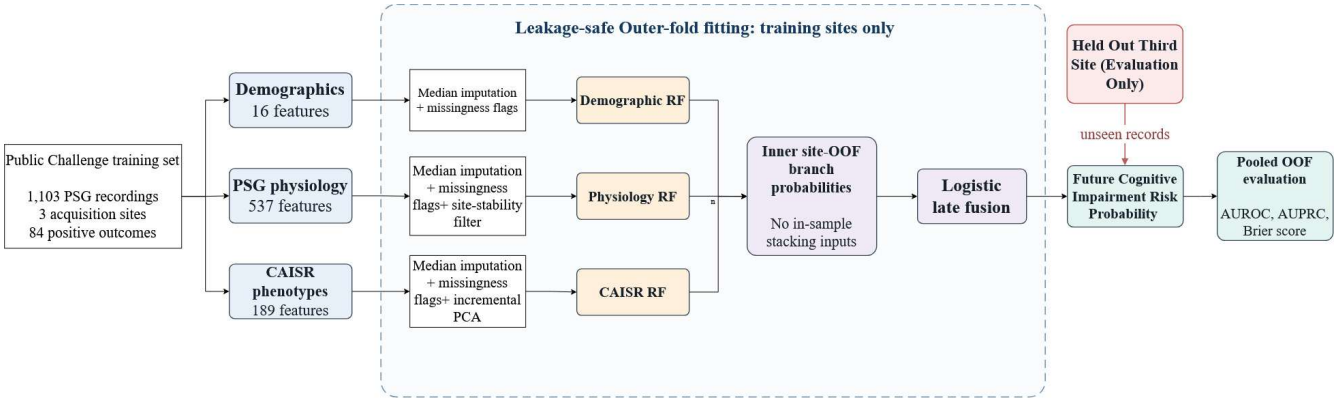

Figure 1. Leakage-safe nested workflow. The outer fold withholds one complete acquisition site. All learned preprocessing and branch models use only the other two sites; the meta-learner receives inner site-held-out out-of-fold branch probabilities. The fitted system then produces risk probabilities for the unseen site.

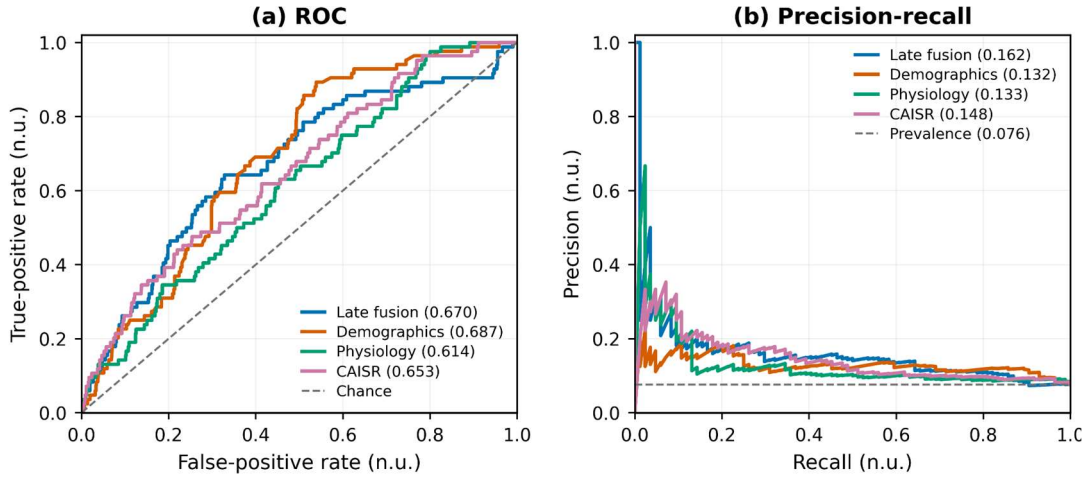

Figure 2. Pooled out-of-fold (a) ROC and (b) precision–recall curves for the late-fusion and three branch models across 1,103 public-training records. Dashed lines denote chance discrimination and the 0.076 outcome prevalence, respectively.

3.3 Physiology site-stability filtering

The physiology branch initially contained 537 candidate variables. The outer folds holding out I0002, I0006, and S0001 retained 319, 324 and 322 variables, respectively, after training-only site-stability filtering. The model subsequently refitted on all 1,103 public-training records retained 238 physiology variables. The full-training feature set was used only for the saved inference artifact and was not substituted into the outer-fold evaluation.

Held-out Site	Records (n)	AUROC	Age-conditioned AUROC
I0002	54	0.731	0.583
I0006	192	0.560	0.415
S0001	857	0.740	0.491
Pooled	**1,103**	**0.670**	**0.473**

Table 2. Late-fusion performance for each held-out acquisition site.

4. Discussion and Conclusion

This study demonstrated a trade-off between global discrimination and rare-positive prioritization. Demographics achieved the highest AUROC (0.687), whereas late fusion achieved the highest AUPRC (0.162) with AUROC 0.670. Compared with the demographic AUPRC of 0.132, fusion improved AUPRC by 22.7% and approximately doubled the prevalence baseline of 0.076. CAISR also outperformed engineered physiology in AUROC (0.653 vs. 0.614) and AUPRC (0.148 vs. 0.133), suggesting complementary information from structured sleep phenotypes, although neither sleep derived branch surpassed demographics in AUROC. These findings align with evidence that conventional sleep-stage measures have inconsistent associations with incident dementia [5], whereas EEG microstructure, spindle–slow-oscillation activity, REM-sleep slowing, brain-age indices and multivariate EEG features show stronger associations with cognitive status or decline

[1,2,6,7,10]. This study extends this evidence through multimodal PSG and automated sleep phenotypes, with its site-held-out evaluation providing a more heterogeneous and conservative generalization estimate than typical single-cohort studies.

Age-conditioned AUROC was 0.473 for fusion, compared with 0.495 for demographics, 0.558 for physiology and 0.573 for CAISR, indicating limited sleep-related information beyond age that was not effectively preserved by the demographic-dominated stacker. Site-level AUROC ranged from 0.560 to 0.740, indicating sensitivity to acquisition setting. The nested design strengthened validity by restricting preprocessing, filtering, model fitting and stacking to training sites and using inner site-held-out probabilities for the meta-learner. However, interpretation remains limited by only 84 positive outcomes, three imbalanced sites, and potential signal-quality or annotation errors. The Brier score of 0.070 was comparable to a prevalence-based predictor and therefore did not establish improved calibration.

Overall, leakage-safe late fusion achieved AUROC 0.670 and AUPRC 0.162 across 1,103 public-training records, improving rare-positive prioritization but showing limited discrimination beyond age and no clear calibration advantage. These are public-training cross-validation estimates, with no official validation score, test score, or ranking; therefore, external or hidden-data evaluation is required before clinical or multicenter generalizability can be claimed. SomnoSense consequently provides a reproducible multicenter evaluation framework and evidence of complementary sleep-related ranking information rather than a clinically validated diagnostic model

Acknowledgments

The authors thank the George B. Moody PhysioNet Challenge organizers and Human Sleep Project contributors for providing the data, scoring framework and reporting guidance. The implementation used open-source Python scientific-computing software. The authors declare no competing interests.

References

[1] Sun H, et al. Machine Learning–Based Sleep Electroencephalographic Brain Age Index and Dementia Risk: An Individual Participant Data Meta-Analysis. JAMA Netw Open. 2026;9(3):e261521. doi: 10.1001/jamanetworkopen.2026.1521.

[2] Haghayegh S, et al. Predicting Future Risk of Developing Cognitive Impairment Using Ambulatory Sleep EEG: Integrating Univariate Analysis and Multivariate Information Theory Approach. J Alzheimers Dis. 2025;108(Suppl 1):S105-S117. doi: 10.1177/13872877251319742.

[3] Wei J, Wang M, Guo Y, Liu Y, Dong X. Sleep Structure Assessed by Objective Measurement in Patients with Mild Cognitive Impairment: A Meta-Analysis. Sleep Med. 2024;113:397-405. doi: 10.1016/j.sleep.2023.12.010.

[4] Yiallourou S, et al. Sleep Architecture and Dementia Risk in Adults: An Analysis of Five Cohorts from the Sleep and Dementia Consortium. Sleep. 2025;48(9):zsaf129. doi: 10.1093/sleep/zsaf129.

[5] Lam AKF, et al. Electroencephalographic Slowing During REM Sleep in Older Adults with Subjective Cognitive Impairment and Mild Cognitive Impairment. Sleep. 2024;47(6):zsae051. doi: 10.1093/sleep/zsae051.

[6] Ujma PP, Bódizs R, Dresler M, Simor P. Multivariate Prediction of Cognitive Performance from the Sleep Electroencephalogram. NeuroImage. 2023;279:120319. doi: 10.1016/j.neuroimage.2023.120319.

[7] Páez A, et al. Sleep Spindles and Slow Oscillations Predict Cognition and Biomarkers of Neurodegeneration in Mild to Moderate Alzheimer's Disease. Alzheimers Dement. 2025;21(2):e14424. doi: 10.1002/alz.14424.

[8] Ganglberger W, et al. Brain Health from Sleep EEG: A Multicohort, Deep Learning Biomarker for Cognition, Disease, and Mortality. NEJM AI. 2026;3(3). doi: 10.1056/AIoa2500487.

[9] Sun H, et al. The Sleep and Wake Electroencephalogram over the Lifespan. Neurobiol Aging. 2023;124:60-70. doi: 10.1016/j.neurobiolaging.2023.01.006.

[10] Leinonen T, et al. Empirical Investigation of Multi-Source Cross-Validation in Clinical ECG Classification. Comput Biol Med. 2024;183:109271. doi: 10.1016/j.compbiomed.2024.109271.

[11] Li Q, Wen S, Sun H, Ganglberger W, Tripathi A, Turley N, Waters S, Gupta A, Gupta A, Ghanta M, Nearing B, Wu H, Stone KL, Robichaux C, Zhang Z, Li Q, Ganjoo G, Silvers CT, Gunapati B, Maski K, Nasiri S, Hwang D, Trotti LM, Katwa U, Clifford GD, Mignot E, Thomas RJ, Westover MB. The Human Sleep Project: A Multi-Center Clinical Polysomnography Dataset Across the Human Lifespan. Sleep. 2026 Aug 5:zsag215. doi: 10.1093/sleep/zsag215. PMID: 42555111.

[12] Nasiri S, et al. CAISR: Achieving Human-Level Performance in Automated Sleep Analysis Across All Clinical Sleep Metrics. Sleep. 2025;48(8):zsaf134. doi: 10.1093/sleep/zsaf134.

[13] Reyna MA, Weigle A, Li Q, Koscova Z, Sun HS, Wen S, Katwa U, Thomas R, Hwang D, Trotti LM, Mignot E, Sameni R, Nasiri S, Westover MB, Clifford GD. Screening for Cognitive Impairment During Sleep Studies: The George B. Moody PhysioNet Challenge 2026. In: Computing in Cardiology 2026, volume 53: 2026. p. 1–4

[14] Collins GS, et al. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. BMJ. 2024;385:e078378. doi: 10.1136/bmj-2023-078378.

[15] Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE. 2015;10(3):e0118432. doi: 10.1371/journal.pone.0118432.

Address for correspondence:
Sejuti Dhali
Khulna University of Engineering & Technology, Khulna-9203, Bangladesh
sejutidhali@gmail.com