Multimodal PSG-Based Prediction of Cognitive Impairment Using an Interpretable XGBoost Framework

Sofia Romagnoli[1,2], Diego Cajal[1,2], Rodrigo Lozano-Puñet[1,2], Josseline Madrid[1,2]

[1] BSICoS Group, I3A, IISA, Universidad de Zaragoza, Zaragoza, Spain
[2] Consorcio Centro de Investigación Biomédica en Red de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Madrid, Spain

Abstract

Polysomnography (PSG) provides a multimodal, non-invasive view of sleep physiology that may contain early markers of cognitive impairment. We developed a physiology-informed XGBoost framework to predict cognitive impairment diagnosis 1–6 years after PSG. A neural-network attention with Grad-CAM was used to identify informative PSG modalities. Along with signal availability and physiological plausibility, this XAI-based analysis guided the selection of 244 features, spanning age, sex, respiratory dynamics, sleep-EEG measures, heart-rate variability, and ECG-derived age. To handle heterogeneous channel availability, seven XGBoost classifiers were trained for all EEG, ECG, and respiration signal combinations. On the 1,103-record development set, five-fold cross-validation yielded an age-conditioned AUROC of 0.543 $\pm$ 0.047, while leave-one-site-out (LOSO) validation yielded 0.520 $\pm$ 0.007. Trained on the 6,600-record set, the model achieved an age-conditioned AUROC of 0.636 and an overall AUROC of 0.837 on the hidden validation set. A revised pipeline improved LOSO performance but obtained lower hidden-validation age-conditioned AUROC, highlighting that external-cohort performance and robustness to site shift are not equivalent. Results emphasize the importance of multisite validation for assessing clinical transportability of PSG-based prognostic models.

1. Introduction

Sleep is closely linked to neurological, cardiovascular, and respiratory health. Polysomnography (PSG) non-invasively captures these systems simultaneously, revealing physiological alterations associated with neurological, cardiovascular, and respiratory dysfunction beyond the sleep disorders for which it is usually acquired [1]. In particular, altered sleep oscillations, autonomic dysregulation, and sleep-disordered breathing have each been associated with cognitive decline [2–4].

Cognitive impairment and mild cognitive impairment often precede dementia and Alzheimer's disease, which remain incurable and carry a heavy societal burden [5, 6]. Because current clinical diagnostics and neuroimaging can be time-consuming and difficult to scale, objective PSG-derived biomarkers may provide an accessible screening tool for early intervention when disease-modifying therapies are most effective [7].

For the 2026 George B. Moody PhysioNet Challenge, we developed an open-source machine learning framework to predict future cognitive impairment from PSG recordings. Our key objectives are: i) to identify predictive physiological patterns in routinely collected sleep data; ii) to develop an interpretable and scalable approach designed to accommodate heterogeneous PSG recordings.

2. Methods

Our approach combined physiology-informed feature engineering with a modality-aware XGBoost framework for cognitive impairment prediction from PSG recordings, which is shown in Figure 1.

2.1. Dataset and preprocessing

The 2026 George B. Moody PhysioNet Challenge dataset was derived from the Human Sleep Project and comprised PSG recordings collected at five U.S. institutions [8, 9]. Recordings included EEG, EOG, EMG, ECG, respiratory effort, airflow, and oxygen saturation, together with sleep annotations. The small training set ($N = 1103$) was used for local model development and cross-validation, while the large training set ($N = 6600$) was used to train official submissions.

To reduce inter- and intra-cohort variability, signals were resampled to a common sampling frequency among modalities, channel availability was assessed, and missing standard bipolar EEG derivations were reconstructed when the corresponding electrodes were available.

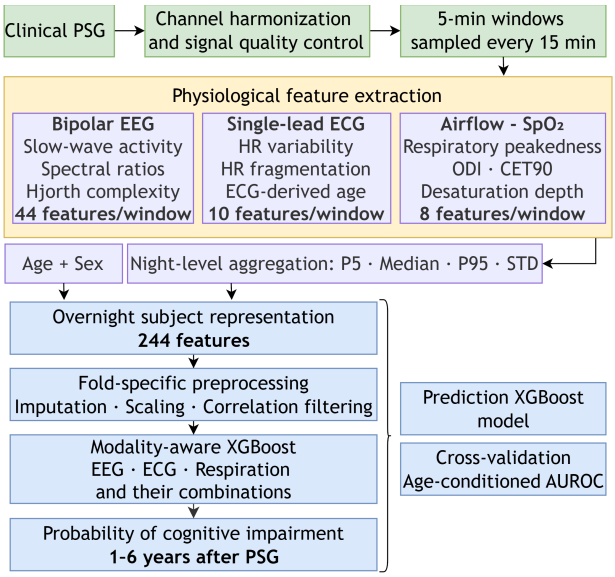

Figure 1. Feature extraction and modality-aware XGBoost framework for cognitive impairment prediction.

Preliminarly, an XAI-based analysis combining neural-network attention and Grad-CAM was used to identify informative PSG modalities. Together with signal availability and physiological plausibility, this analysis guided the selection of EEG, ECG-derived measures, and respiratory signals for the subsequent XGBoost framework. This XAI analysis was used as an exploratory development step and was not part of the final inference pipeline.

Physiological features were extracted from 5-min windows sampled every 15 min throughout each recording, as detailed in Sections 2.1.1, 2.1.2, and 2.1.3. For each segment-level feature, except ECG-derived age, subject-level representations comprised the median, 5th and 95th percentiles, and standard deviation across windows. ECG-derived age was summarized by its median, and the difference between ECG-derived and chronological age was included as feature. Together with chronological age and sex, this produced a fixed 244-feature raw representation.

Missing continuous values were imputed using a five-nearest-neighbor imputer, whereas missing categorical values were replaced by the mode. Continuous features were standardized and redundant continuous variables were removed using a Pearson-correlation threshold of $|r| > 0.8$. All data-dependent preprocessing steps were fitted on the corresponding training partition during cross-validation and applied unchanged to its validation partition.

2.1.1. EEG features

Bipolar EEG derivations C3–M2, C4–M1, F3–M2, and F4–M1 were filtered with a fourth-order Butterworth band-pass between 0.3 and 35 Hz, z-score standardized, and segmented into 30-s epochs. EEG features were extracted separately for each derivation and included five slow-wave activity measures (total slow-wave count, slow-wave density [10, 11], mean peak-to-peak amplitude, mean negative slope, and mean duration), three spectral slowing measures (relative delta power, theta-to-alpha power ratio, and theta-to-beta power ratio), relative beta power, Hjorth complexity, and delta-power variability.

2.1.2. ECG and heart rate features

ECG recordings were resampled to 200 Hz, and band-pass filtered between 0.5 and 50 Hz after removing power-line interference with a 59.5–60.5 Hz notch filter. Recordings containing NaN values or more than 20% zero-valued samples were discarded. An ECG-derived age estimate was obtained from the preprocessed waveform using the PROPHECG-Age-Single 1D ResNet model [12].

R-peaks were detected using the Pan–Tompkins algorithm [13], and NN intervals were corrected for ectopic beats using a local median-based procedure; the percentage of corrected intervals was retained as the ECTOPIC feature. The corrected NN series was interpolated using second-order PCHIP before heart-rate variability (HRV) analysis. Time-domain HRV features included the average NN interval (AVNN), the standard deviation of NN intervals (SDNN), and the root mean square of successive NN interval differences (RMSSD), while high-frequency power (HF) was estimated over 0.15–0.40 Hz using a Lomb–Scargle periodogram [14]. HR fragmentation was characterized by classifying consecutive NN differences as acceleration, deceleration, or no-change, and grouped into accelerative or decelerative segments. Three fragmentation indices were computed: percentage of inflection points (PIP), percentage of NN differences in long segments (PNNLS), and percentage of NN differences in short segments (PNNSS) [14].

2.1.3. Respiratory features

Respiratory signals were resampled to 25 Hz, and a peakedness measure [15] was used to quantify the concentration and stability of spectral energy within the expected respiratory band for abdominal and thoracic efforts, nasal pressure, and airflow. When several channels represented the same group, the estimate with the best available quality measure was retained. For oxygen saturation (SpO_2), the extracted metrics were mean SpO_2, cumulative time with $SpO_2 < 90\%$ (CET90), oxygen desaturation index (ODI), and mean desaturation depth (ODI-deepness).

2.2. Modality-aware XGBoost ensemble

To handle heterogeneous PSG modality availability, we designed a modality-aware ensemble comprising seven XGBoost classifiers: three unimodal, three bimodal, and one multimodal model using EEG, ECG, and respiratory features. Age and sex were included in all models, and class imbalance addressed using class-weighted XGBoost.

Hyperparameters were optimized using age-conditioned AUROC as the search score. The development pipeline used five-fold stratified nested cross-validation with fold-specific preprocessing. Hyperparameters selected across the outer folds were combined into consensus values and used to fit the final ensemble on the complete training set. For each recording, the classifier matching the available PSG modalities was selected dynamically, avoiding to impute an entirely missing modality. Binary predictions were obtained using a probability threshold optimized on out-of-fold predictions by maximizing the F1-score.

To assess robustness to site-specific distribution shifts independently of model selection, we additionally performed leave-one-site-out (LOSO) nested cross-validation on the small training set, holding out I0002, I0006, and S0001 in turn. This analysis was used for reporting only and did not modify the model or hyperparameters used for the selected Challenge pipeline.

2.3. Model-development comparison

During subsequent model development, we evaluated a revised pipeline incorporating a different physiological feature representation, modified temporal aggregation, and a redesigned modality-routing strategy. Direct respiratory features were removed as standalone predictors, while respiration-informed HRV measures were incorporated using Biosigpy [16]. This revised pipeline was retained as a secondary comparison because it showed substantially stronger LOSO performance across the available training sites, although it achieved a lower age-conditioned AUROC and Reward on the official hidden validation set.

3. Results

Internal validation results are summarized in Table 1, and official hidden-validation results are shown in Table 2.

4. Discussion

We designed a multimodal PSG framework to preserve an interpretable relationship between physiology and prediction while accommodating heterogeneous clinical recordings. Modality selection was informed by a preliminary XAI analysis, whereas the final predictor relied on explicit physiology-derived EEG, ECG, and respiratory features. This design was motivated by the substantial variability of clinical PSG across acquisition equipment, channel configurations, sampling frequencies, and populations.

A central finding was that model ranking depended strongly on the validation domain. The selected model achieved the highest age-conditioned AUROC, AUROC, Reward and F-measure on the official hidden validation cohort (see Table 2), whereas the revised pipeline, based on a different feature representation and modality-routing strategy, consistently performed better when individual training sites were held out (see Table 1). Its LOSO age-conditioned AUROC improved from 0.520 ± 0.007 to 0.626 ± 0.076, with higher performance for each of I0002, I0006, and S0001 sites. Thus, performance on a particular external cohort and robustness to site shift were related but non-equivalent properties. Since several processing and modeling components changed simultaneously, these results should not be attributed to feature-space reduction alone. Internal cross-validation was performed on the 1,103-record development set, whereas official submissions were trained on 6,600 records, so absolute scores between both settings should not be interpreted as directly interchangeable. From a translational perspective, the stronger LOSO performance of the revised pipeline provides greater evidence of robustness to an unseen acquisition site. Conversely, its lower performance on the official validation cohort illustrates that cross-site validation over a limited number of training domains cannot guaranty performance in every new population.

The difference between the overall AUROC (0.837) and teh age-conditioned AUROC (0.636) suggests that model discrimination may rely on age confounding caused by training data imbalances. To prevent age from dominating predictions, ECG-derived biological age and the biological-chronological age gap were incorporated alongside chronological age as inputs. This demographic sensitivity aligns with literature reporting inconsistent physiological signatures of cognitive decline across age groups [3, 17]. Furthermore, sample weighting within the XGBoost loss function proved insufficient to resolve class imbalance, suggesting that targeted data augmentation should be prioritized in future work.

To balance computational efficiency with temporal physiological dynamics, features were extracted using 5-min sliding windows every 15 min and summarized across recordings using central tendency and dispersion metrics. While a stage-specific selection strategy using CAISR automated sleep staging was tested, it offered no performance gains in local validation. Therefore, an annotation-independent sampling approach was chosen to reduce computational complexity and avoid reliance on automated staging models without sacrificing predictive power.

Table 1. Internal age-conditioned AUROC on the 1,103-record development set. OOF denotes pooled out-of-fold.

Model	5-fold mean	5-fold OOF	I0002	I0006	S0001	LOSO mean	LOSO OOF
Selected Challenge model	0.543 ± 0.047	0.549	0.521	0.528	0.511	0.520 ± 0.007	0.531
Revised pipeline	0.580 ± 0.075	0.555	0.729	0.601	0.548	0.626 ± 0.076	0.555

Table 2. Performance of the two evaluated pipelines on the official hidden validation set.

Model	Reward	Age-cond. AUROC	Age-weighted AUROC	AUROC	AUPRC	Accuracy	F-measure
Selected Challenge model	0.080	0.636	0.597	0.837	0.332	0.893	0.377
Revised pipeline	0.016	0.614	0.558	0.828	0.332	0.936	0.099

5. Conclusion

A physiology-informed, modality-aware XGBoost framework has been developed for predicting future cognitive impairment from heterogeneous PSG recordings. The divergence between overall and age-conditioned performance highlights demographic confounding, emphasizing the need for targeted strategies to address age imbalances. Additionally, discrepancies between official validation and LOSO results demonstrate that single-cohort accuracy and cross-site generalization are not equivalent. Extensive multicenter validation remains critical to ensure translational robustness across unseen clinical sites.

Acknowledgments

Funding was provided by EU MSCA COFOUND postdoct project ARISTOS, GA:101081334; by Ministerio de Ciencia, Innovación y Universidades (MCIN)/Agencia Estatal de Investigación (AEI)/10.13039/501100011033 under Grant PID2024-160041OB-I00; by the Gobierno de Aragon (Reference Group BSICoS T39-23R); and by CIBER-BBN through Instituto de Salud Carlos III. The computation was performed at the High Performance computing platform of the NANBIOSIS ICTS.

References

[1] Miller MA. The role of sleep and sleep disorders in the development, diagnosis, and management of neurocognitive disorders. Front Neurol Oct. 2015;6:224.

[2] Mander BA. Local sleep and Alzheimer's disease pathophysiology. Front Neurosci Sep. 2020;14:525970.

[3] Forte G, Favieri F, Casagrande M. Heart rate variability and cognitive function: A systematic review. Front Neurosci Jul. 2019;13:710.

[4] Leng Y, et al. Association of sleep-disordered breathing with cognitive function and risk of cognitive impairment: A systematic review and meta-analysis. JAMA Neurology Oct. 2017;74(10):1237–1245.

[5] Dhakal A, Bodrin B. Cognitive deficits. StatPearls Treasure Island FL StatPearls Publishing Feb. 2023; PMID:32644478.

[6] Lü W, Zhang M, Yu W. Differentiating Alzheimer's disease from mild cognitive impairment: a quick screening tool based on machine learning. BMJ Open Dec. 2023; 13:e073011.

[7] Antonioni A, et al. Sleep disturbances across dementias and cognitive decline: Study protocol for a systematic review and network meta-analysis of polysomnographic findings. J Clin Med Oct. 2025;14:7437.

[8] Sun H, et al. The human sleep project. Brain Data Science Platform, version 2.0, 2023.

[9] Screening for cognitive impairment during sleep studies: The George B. Moody Physionet Challenge 2026. https://moody-challenge.physionet.org/2026/.

[10] Mensen A, Riedner B, Tononi G. Sleep wave analysis toolbox for matlab. Zenodo, 2015. URL https://doi.org/10.5281/zenodo.14985. MATLAB software.

[11] Mensen A, Riedner B, Tononi G. Optimizing detection and analysis of slow waves in sleep EEG. J Neurosci Methods Dec. 2016;274:1–12.

[12] Park S, et al. Wearable device derived electrocardiographic age and its association with atrial fibrillation. npj Digit Med 2026;9:157.

[13] Joge SD. QRS detection using Pan–Tompkins algorithm from ECG signal. MATLAB Central File Exchange, 2026. MathWorks.

[14] Costa MD, et al. Prediction of cognitive decline using heart rate fragmentation analysis: The multi-ethnic study of atherosclerosis. Front Aging Neurosci 2021;13:708130.

[15] Kontaxis S, et al. ECG-derived respiratory rate in atrial fibrillation. IEEE TBME 2019;67(3):905–914.

[16] Cajal D, Armañac-Julián P. Biosigpy, August 2026. URL https://doi.org/10.5281/zenodo.22112728.

[17] Derby CA, et al. Heart rate variability, cognitive performance and cognitive impairment in older adults: Cross-sectional results from the einstein aging study. Health Sci Rep Aug. 2025;8:e71152.

Address for correspondence:

Sofia Romagnoli

C/ Mariano Esquillor Gómez, s/n, 50018 Zaragoza

sromagnoli@unizar.es