Age-Residualized Physiological Biomarkers and Gradient-Boosted Trees for Predicting Cognitive Impairment from Polysomnography

Paul Fischer[1], Simon Boehi[1], Sidaty El Hadramy[1], Volodimir Buchakchiyskiy[1], Hélène Corbaz[1], Max Krähenmann[1], Philippe C. Cattin[1], Ece Ozkan[1]

[1] Department of Biomedical Engineering, University of Basel, Switzerland

Abstract

Cognitive impairment is associated with disrupted sleep microarchitecture, including reduced slow-wave activity, altered spindle density, impaired spindle-slow-oscillation coupling, autonomic dysregulation, and intermittent hypoxia. We present our entry to the PhysioNet/Computing in Cardiology Challenge 2026, which asks participants to predict a future diagnosis of cognitive impairment (mild cognitive impairment, Alzheimer's disease, or dementia) from a single overnight polysomnography (PSG) recording. Rather than learning directly from raw multi-channel signals, we extract a bank of 136 *literature-motivated physiological features from the measurements, stratified by sleep stage. We further compute an* age-residual *feature for a curated set of* 23 *biomarkers with established age trends. The resulting tabular representation is classified with a bagging ensemble of 15 gradient-boosted trees (XGBoost), with hyperparameters selected by randomized search under stratified cross-validation. On the official challenge validation set, our submitted model trained on the large version of the dataset reaches an age-conditioned AUROC of* 0.669.

1. Introduction

Sleep is an essential physiological process and quantifying its quality can provide clinically relevant markers for a wide range of diseases, including cognitive impairment and other chronic conditions [1]. One of the most widely used methods for sleep analysis is polysomnography (PSG). PSG recordings are multi-modal and complex, capturing physiological activity from multiple systems (cerebral, cardiovascular, and muscular function) throughout sleep. As a result, PSG features may reveal underlying or emerging pathological processes, including neurological, cardiovascular, and respiratory conditions, potentially in their early stages [2].

Cognitive impairment in particular has been linked to specific, well-characterized changes in sleep microarchitecture: reduced slow-wave activity and flattened slow-wave morphology, decreased sleep-spindle density and altered spindle-slow-oscillation coupling, cardiovascular autonomic dysregulation during sleep, and intermittent hypoxia from sleep-disordered breathing [3, 4]. Machine learning models can capture complex patterns in multi-modal biomedical data and may support clinical decision-making from sleep recordings. In this work, we develop and evaluate a model to predict future cognitive impairment diagnoses from PSG data, aiming to provide predictions that help clinicians identify patients who may be more likely to develop cognitive impairment.

Rather than treating the task as an end-to-end deep-learning problem on raw multi-channel PSG, we follow a feature-engineering pipeline: we compute a set of interpretable, literature-motivated physiological biomarkers [3, 5] per recording and train a tree-based classifier on top of them. We further introduce an age-residualization step to account for age-specific trends within features. The objective is to predict future cognitive impairment from PSG recordings using machine learning. We hypothesize that PSG patterns contain information predictive of subsequent cognitive impairment. This paper presents our participation in the PhysioNet Challenge 2026.

2. Method

We develop a binary machine learning model to predict future cognitive impairment in patients grouped as mild cognitive impairment, Alzheimer's disease, or dementia. This study is retrospective, using existing PSG data and associated clinical labels. Our approach is visualized in Figure 1.

2.1. Data

The data are obtained from the Human Sleep Project database [6]. This multicenter, anonymized dataset includes participants who underwent polysomnography (PSG). Sleep stages and related measurements were generated using the Complete AI Sleep Report (CAISR) frame-

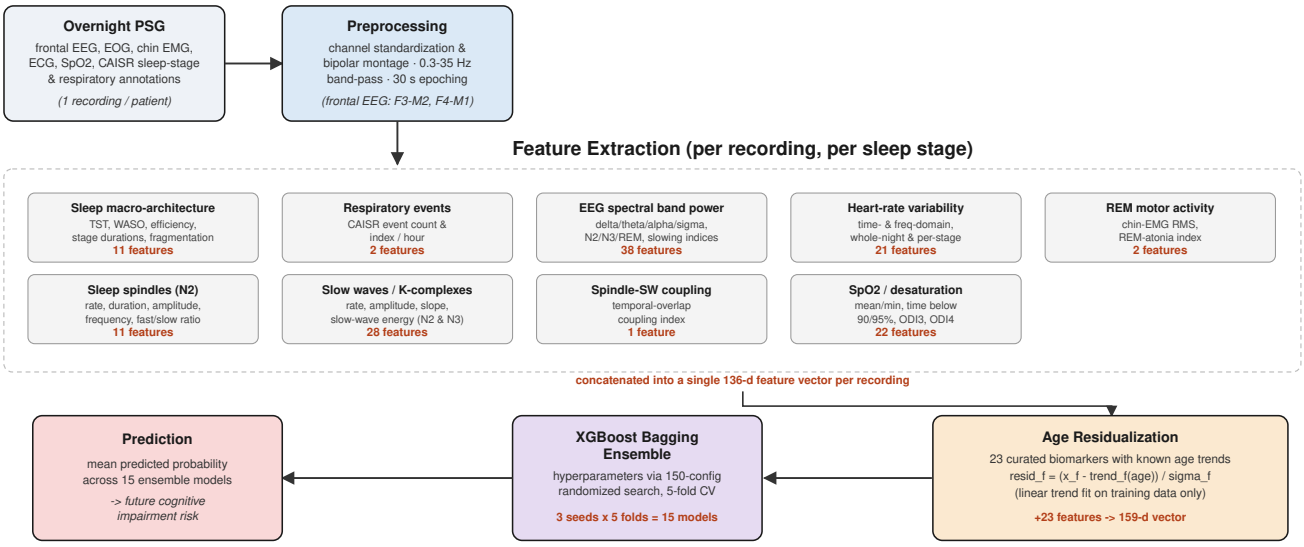

Figure 1. Overview of the feature-extraction and classification pipeline.

work [7], and sleep scoring was performed according to the American Academy of Sleep Medicine (AASM) guidelines [8]. The dataset includes recordings from five centers: three centers were used for training (1090 annotated cases), one for validation (XXX cases), and one for testing (XXX cases).

For model development, we used the locally released training data ($n = 1090$ recordings across three centers that had associated algorithmic annotations), for which the target label, *Cognitive Impairment*, is severely imbalanced (84 positive recordings, $\approx 8\%$). Because a random patient-level split would test interpolation within centers the model already trains on rather than generalization to an unseen center, as required by the Challenge, we held out one entire center as an internal proxy for the external validation site and split the remaining two centers $85\%/15\%$, stratified by label, into an internal training set ($n = 765$) and an internal validation set used for model and hyperparameter selection ($n = 135$).

2.2. Model

2.2.1. Preprocessing

Raw channel labels are highly inconsistent across recording sites and are first mapped to a canonical naming scheme via an alias table. Standard referential PSG derivations (frontal, central, occipital EEG; EOG; chin EMG; leg leads), each referenced to the contralateral mastoid, are constructed by subtraction where both electrodes are available. All EEG channels used for feature extraction are zero-phase band-pass filtered (0.3-35 Hz, 4th-order Butterworth). Because the montage is incomplete for one training center, we restrict all EEG-derived features to the two frontal derivations (`F3-M2`, `F4-M1`), which are available at every site and for which frontal slow-wave and K-complex activity are among the best-characterized cognitive-impairment markers in the sleep literature [3, 5]. The CAISR sleep-stage annotation is converted to 30-second epochs (the standard AASM scoring epoch) by majority vote over the annotated samples in each epoch.

2.2.2. Feature Extraction

For each recording, we compute 136 scalar features across nine families, most of them stratified by sleep stage (Table 1): (i) whole-night sleep macro-architecture (total sleep time, sleep efficiency, wake after sleep onset, stage durations, awakenings, fragmentation index); (ii) respiratory-event count and index derived from the CAISR respiratory annotation; (iii) EEG spectral band power (delta, theta, alpha, sigma) per frontal channel, estimated by Welch's method and averaged within N2, N3, and REM epochs, plus derived slowing (delta/theta) and relative-sigma indices; (iv) sleep-spindle rate, duration, amplitude, frequency, and fast/slow ratio in N2, detected with YASA [9]; (v) slow-wave and K-complex rate, duration, amplitude, slope, and slow-wave-energy burden, detected separately in N2 (K-complexes) and N3 (slow-wave sleep); (vi) a temporal-overlap spindle-slow-wave coupling index; (vii) heart-rate variability (time- and frequency-domain, whole-night and per-stage), computed from ECG R-peaks detected with NeuroKit2 [10]; (viii) SpO_2 distributional statistics, time below clinical desaturation thresholds, and 3%/4% Oxygen Desaturation Index, whole-night and per stage; and (ix) a chin-EMG-based

REM-atonia index (REM-epoch RMS relative to stable-NREM RMS), a simplified proxy for REM Sleep Without Atonia, itself a strong prodromal marker of neurodegeneration [11](full details in our code repository).

Table 1. Number of raw features per family.

Feature family	Count
Sleep macro-architecture	11
Respiratory events	2
EEG spectral band power	38
Sleep spindles	11
Slow waves / K-complexes	28
Spindle-slow-wave coupling	1
Heart-rate variability	21
SpO_2 / desaturation	22
REM motor activity (chin EMG)	2
Total	**136**

2.2.3. Age-Residualized Features

The Challenge's primary metric, age-conditioned AUROC, compares a case and a control only when their ages differ by at most two years, so only the age-independent deviation of a biomarker is informative for that comparison. We therefore select 23 features with well-established age trends in the sleep-medicine literature (e.g., total sleep time, N3 duration, slow-wave slope and energy, spindle density and amplitude, HRV indices, the REM-atonia index, and desaturation burden) and, for each, fit a linear age trend $\hat{x}_f(\text{age}) = \alpha_f \cdot \text{age} + \beta_f$ by ordinary least squares on the training data only. The corresponding feature is the standardized residual,

$$\texttt{resid}_f = \frac{x_f - \hat{x}_f(\text{age})}{\sigma_f}, \tag{1}$$

where σ_f is the training-residual standard deviation. The trend is fit once on training data and only applied to validation, test, or inference records, to avoid leaking age-outcome relationships from held-out data. This adds 23 columns, for a total feature dimensionality of 159.

2.2.4. Classifier

The 159-dimensional feature vector is passed directly to an XGBoost classifier [12] (no scaling, since tree splits are scale-invariant, and no explicit imputation, since XGBoost learns a default split direction for missing values natively). Hyperparameters (number of trees, tree depth, learning rate, row/column subsampling, minimum split-loss reduction, L1/L2 regularization, minimum child weight, and a class-imbalance reweighting term) are selected by randomized search (150 sampled configurations) under 5-fold stratified cross-validation, scored by ROC AUC. The final model is a repeated k-fold bagging ensemble: three independent 5-fold stratified partitions of the training data yield 15 XGBoost models, each trained on one fold's $\sim$80% training portion with the selected hyperparameters and a distinct random seed. At inference, the predicted probability is the mean of the 15 models' outputs.

2.3. Training and Evaluation

We report results on the official Challenge validation set, evaluated by the organizers against the seven official Challenge metrics using the organizers' evaluation script. We compare our final submitted approach (Model 3: XGBoost with age-residual features and bagging, submission 2626) against two ablations submitted separately: (i) Model 1, XGBoost trained on the raw features only, without age residualization or bagging (submission 2453); and (ii) Model 2, TabICL, a tabular in-context-learning foundation model [13], trained on the identical raw feature matrix, as an architecturally different baseline that is not part of our final submission (submission 2272) All models were fitted on the large version of the challenge dataset.

3. Results

Table 2. Official Challenge metrics on the official challenge validation set. Model 1 (submission 2453): XGBoost on raw features. Model 2 (submission 2272): TabICL on raw features. Model 3 (submitted, submission 2626): XGBoost with age-residual features and bagging.

Metric	Model 1	Model 2	Model 3
Reward	0.006	0.08	0.007
Age-cond. AUROC	0.685	0.635	0.669
Age-weighted AUROC	0.657	0.573	0.635
AUROC	0.764	0.727	0.802
AUPRC	0.212	0.161	0.297
Accuracy	0.932	0.935	0.932
F-measure	0.020	0.000	0.040

On the official validation set, the three approaches diverge depending on the metric considered. Model 1 (plain features, no age residualization or bagging) achieves the highest age-conditioned AUROC (0.685), the metric the Challenge ranks submissions on, narrowly ahead of our submitted Model 3 (0.669). Model 2 (TabICL) trails both on this metric (0.635). The ordering reverses on plain AUROC and AUPRC, where Model 3 is highest (AUROC 0.802, AUPRC 0.297), followed by Model 1 (AUROC 0.764, AUPRC 0.212) and Model 2 (AUROC 0.727, AUPRC 0.136). Reward and F-measure are low for all three models (Reward 0.006–0.008; F-measure 0.000–0.040) at the fixed 0.5 decision threshold and $\approx$8% pos-

itive prevalence in the training data, consistent with predicted probabilities rarely crossing 0.5 for any of the three approaches regardless of the class-imbalance reweighting term (`scale_pos_weight`) selected during randomized search. Because AUROC-type metrics, including the Challenge's primary age-conditioned AUROC, are invariant to any monotonic rescaling of the predicted score, this threshold sensitivity does not by itself favor or disfavor any of the three approaches on the primary metric, but it is a practically important limitation of a fixed 0.5 cutoff under severe class imbalance.

4. Discussion and Conclusion

We presented a feature-engineering and gradient-boosted-tree approach to predicting future cognitive impairment from overnight PSG for the PhysioNet/CinC Challenge 2026. Our pipeline computes 136 literature-grounded physiological features. On the official Challenge validation set, our submitted approach (Model 3) reached an age-conditioned AUROC of 0.669, close to but not exceeding a plain-feature baseline without age residualization or bagging (Model 1, 0.685), while outperforming both baselines on plain AUROC and AUPRC. A tabular foundation-model baseline (Model 2, TabICL) trailed both gradient-boosted-tree variants on every metric except age-weighted AUROC, where all three models were close. Because the official validation numbers reflect a single evaluation on the organizers' held-out set, we cannot directly quantify their sampling uncertainty. During development we estimated, on our internal, site-disjoint holdout, a bootstrap 95% confidence interval on age-conditioned AUROC of roughly $[0.62, 0.84]$ for a comparably sized split, which suggests that differences of a few points between models of this kind should be interpreted cautiously even though it is not a formal uncertainty estimate for the official table itself.

Given the limited number of positive training examples ($n = 84$) relative to the feature dimensionality, results should be interpreted with this uncertainty in mind, and repeated-split or cross-site evaluation would give a more reliable estimate of generalization than a single validation-set evaluation. Restricting EEG features to frontal derivations, to accommodate one center's incomplete montage, and using a coarse temporal-overlap rather than phase-based spindle-slow-wave coupling index, are further simplifications that could be revisited in future iterations. Overall, our results support the feasibility of a compact, interpretable, feature-based approach to this task, while highlighting the importance of aligning model design, and threshold choice in particular, with the specific metrics used for evaluation.

Acknowledgments

The authors would like to thank Prof. Dr. Alexandre Datta and Lisa Cazzato from the Children Hospital of Basel (UKBB) for their help in defining clinically relevant features and feedback on their approach. We would like to thank SCIcore for providing access to the GPU clusters used for model training.

References

[1] Ramar K et al. Sleep is essential to health: an american academy of sleep medicine position statement. Journal of Clinical Sleep Medicine October 2021;17(10):2115–2119.

[2] Miller MA. The role of sleep and sleep disorders in the development, diagnosis, and management of neurocognitive disorders. Frontiers in Neurology 2015;6.

[3] Mander BA et al. Sleep: a novel mechanistic pathway, biomarker, and treatment target in the pathology of alzheimer's disease? Trends in neurosciences 2016; 39(8):552–566.

[4] Winer JR et al. Sleep as a potential biomarker of tau and β-amyloid burden in the human brain. The Journal of Neuroscience 2019;39(32):6315–6324.

[5] Mander BA et al. β-amyloid disrupts human nrem slow waves and related hippocampus-dependent memory consolidation. Nature neuroscience 2015;18(7):1051–1057.

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

[7] Nasiri S et al. Caisr: achieving human-level performance in automated sleep analysis across all clinical sleep metrics. Sleep August 2025;48(8):zsaf134.

[8] Grigg-Damberger MM. The aasm scoring manual four years later. Journal of Clinical Sleep Medicine 2012; 8(3):323–332.

[9] Vallat R, Walker MP. An open-source, high-performance tool for automated sleep staging. elife 2021;10:e70092.

[10] Makowski D et al. Neurokit2: A python toolbox for neurophysiological signal processing. Behavior research methods 2021;53(4):1689–1696.

[11] Postuma RB, Berg D. Advances in markers of prodromal parkinson disease. Nature Reviews Neurology 2016; 12(11):622–634.

[12] 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.

[13] Qu J et al. Tabicl: A tabular foundation model for in-context learning on large data. arXiv preprint arXiv250205564 2025;.

Address for correspondence:

Paul Fischer

Hegenheimermattweg 167b, 4123 Allschwil, Schweiz

paul.fischer@unibas.ch