Prediction of Cognitive Impairment from Polysomnography Using XGBoost and a Pretrained Tabular Foundation Model (TabPFN V2)

Elaine Li

College of Arts and Sciences, Emory University, Atlanta, United States of America

Abstract

As part of the George B. Moody PhysioNet Challenge 2026, our team, insomniac111, developed two open-source machine-learning pipelines to predict future cognitive impairment from polysomnography (PSG). The public data included a 1,103-record small training set and a 6,600-record large set from three institutions, with heterogeneous physiological signals, Complete AI Sleep Report (CAISR) annotations, human annotations, demographics, and diagnostic labels. After channel harmonization and 200-Hz resampling of EEG, EOG, and ECG, XGBoost used a compact 156-feature representation: 12 demographic variables, 105 physiological summaries, 31 CAISR measures, and eight missingness indicators. TabPFN V2 used an expanded, quantile-normalized table of 243 features that additionally included regional EEG spectra, heart-rate variability, enhanced oxygenation, human annotations, and stage-conditional physiology. We used leave-one-site-out validation to estimate cross-source generalization. On the blinded evaluation, small-set trained XGBoost achieved an age-conditioned AUROC of 0.765, large-set XGBoost achieved 0.716, and small-set TabPFN V2 achieved 0.645. Notably, large-set XGBoost slightly improved conventional AUROC over small-set XGBoost (0.855 versus 0.851) while reducing the official score. This divergence suggests that additional data strengthened global discrimination but not ranking among age-matched patients. Possible causes include greater reliance on age or site-associated signals, training-to-evaluation shift, and hyperparameter or objective mismatch. These findings favor metric-aligned sampling, feature-matched ablations, and site-aware blending over an assumption that increasing training-set size alone will improve Challenge performance.

1. Introduction

Sleep is tightly coupled to brain health, and disturbances in sleep architecture, oxygenation, autonomic regulation, and arousal may precede clinically recognized neurocognitive disease [1,2]. The 2026 George B. Moody PhysioNet Challenge invited teams to develop open-source algorithmic approaches for using polysomnography (PSG), which records various physiological signals during sleep studies, to predict future diagnoses of cognitive impairment: given one PSG and basic demographics, an algorithm must estimate whether the patient will receive at least two qualifying diagnoses of mild cognitive impairment, Alzheimer disease, or dementia, with the first diagnosis occurring 1-6 years after the PSG and the diagnoses separated by at least one week [1].

The provided training data are a curated subset of the Human Sleep Project (HSP), a multi-institution clinical resource designed to preserve real-world variation in patients, equipment, montages, sampling rates, and signal quality [3]. The public training cohorts come from sites S0001, I0002, and I0006. The validation set is hidden so that the organizer can evaluate teams on it during the Challenge. The ranking metric is age-conditioned AUROC: a positive patient should receive a higher probability than a negative patient within a two-year age neighborhood. A prevalence-based reward evaluates binary decisions, while conventional AUROC, AUPRC, accuracy, and F-measure are secondary measures [1].

Raw PSG permits expressive end-to-end models, including modern convolutional networks such as ConvNeXt V2 [11], but each record contains an entire heterogeneous night and the public sets contain only 1,103 or 6,600 labeled recordings. We therefore compared two tabular strategies. XGBoost learns regularized ensembles of decision trees directly from a compact engineered table [10]. TabPFN V2 transfers a transformer-based inference algorithm learned from millions of synthetic tabular tasks [7,8] and was supplied a broader feature representation. The blinded comparison evaluates the complete pipelines rather than a feature-matched contest between classifiers.

2. Methods

2.1 Training Data and endpoint

The primary development analysis used the 1,103-record small training set from sites S0001, I0002, and I0006; we subsequently repeated the XGBoost pipeline on the 6,600-record large set from the same sources. Each record could provide EDF physiological signals, CAISR outputs, technician annotations, demographics, and a training-only diagnostic label [1,4]. Sessions were linked by BidsFolder and SessionID and retained when the binary endpoint and required inputs were available. TabPFN V2 was evaluated on the small set. The hidden validation (I0004) and test (I0007) cohorts originate from different sources, making both transportability and training-cohort composition central design considerations.

2.2 Harmonization and feature extraction

Channel labels were standardized with the official channel table, and referential signals were converted to conventional bipolar derivations when possible. EEG,

EOG, and ECG were resampled to 200 Hz with polyphase anti-alias filtering; the XGBoost path interpolated isolated non-finite samples before filtering. Both extractors excluded invalid values before summary calculation and limited extremes by robust percentile clipping. Each selected signal was summarized by 15 distributional and dynamical descriptors: mean, standard deviation, five percentiles, interquartile range, root-mean-square amplitude, zero-crossing rate, line length, Hjorth mobility and complexity [13], a modality-specific descriptor, and recording duration.

The XGBoost representation deliberately remained compact. It contained 12 demographic features; 105 summaries from one available channel in each of seven groups (EEG, EOG, chin EMG, leg EMG, ECG, respiration, and SpO2); seven channel missingness flags; 31 CAISR features describing sleep composition, event burden, timing, bouts, and confidence; and one CAISR missingness flag. XGBoost preserved the resampled signals' amplitude scale. Human-annotation availability was audited but human-derived values were not used, yielding 156 features.

Table 1. Patient-level feature representation.

Feature block	XGBoost	TabPFN V2
Demographics	12	12
Signal summaries + flags	112 (7 groups)	144 (9 groups)
Spectral / HRV / SpO2 extensions	0	29
CAISR + missing flag	32	37
Human annotations + flag	0	13
Stage-conditional physiology	0	8
Total	**156**	**243**

TabPFN used 243 features. It separated frontal, central, and occipital EEG; 200Hz resampled and z-score normalized each EEG, EOG, and ECG lead; added regional spectral powers, five HRV measures, six SpO2 measures, five respiratory-event subtype indices, 12 human annotation summaries, and eight stage-conditional cross-modal measures. Missing blocks were zero-filled and flagged. A QuantileTransformer mapped the completed training table to a normal marginal distribution before TabPFN fitting.

2.3 XGBoost classifier

XGBoost constructs an additive ensemble of shallow decision trees [10]. At each boosting round, a new tree is selected using first- and second-order derivatives of the binary logistic loss, so it corrects errors made by the preceding ensemble. Tree depth, shrinkage, minimum child weight, row and feature subsampling, split penalties, and L2 regularization constrain variance. This piecewise-constant structure is well suited to clinical thresholds and nonlinear interactions, does not require feature scaling, and naturally accommodates mixed magnitudes among demographic, signal, and annotation variables.

We used histogram-based XGBClassifier models with 250-500 trees, maximum depth 3-4, learning rate 0.03-0.05, row subsampling 0.80-0.85, feature subsampling 0.75-0.85, and increasing child-weight and regularization constraints. Three prespecified configurations were compared by mean leave-one-site-out AUROC. The selected model was refitted on all eligible records; its continuous positive-class probability was submitted, while an out-of-fold F1-optimized threshold supplied the required binary output.

2.4 TabPFN V2 classifier

TabPFN is a Prior-data Fitted Network: a transformer is pretrained across millions of synthetic tables drawn from a broad prior over data-generating mechanisms [7,8]. For a new task, labeled training rows provide context and unlabeled query rows receive posterior-predictive class probabilities without gradient-based retraining of the full network. TabPFN V2 represents individual cells and alternates attention across features within a row and samples within a column, making it invariant to row and column order and robust to many small-table irregularities [8]. Our offline implementation used a packaged V2 classifier checkpoint with 32 ensemble estimators. The fitted context, quantile transformer, feature dimension, cross-validation summaries, and decision threshold were saved for deterministic inference.

2.5 Site-aware validation and scoring

For both pipelines, each training source (S0001, I0002, and I0006) was held out once while the other two supplied training data. Out-of-fold probabilities were pooled for AUROC, AUPRC, and threshold estimation. XGBoost used the mean site AUROC to select its hyperparameters; TabPFN evaluated one fixed configuration. Final small-set models were fitted on 1,103 records, and the large-set XGBoost experiment repeated the extraction and model-selection pipeline on 6,600 records. This design is harder and more deployment-relevant than a random split because site-specific acquisition patterns cannot be shared between training and validation rows.

Internal model selection nevertheless used conventional AUROC, whereas the Challenge score ranks a positive above a negative only when their ages differ by no more than two years [1]. Consequently, future development should optimize the official age-conditioned scorer directly. The reported blinded scores measure ranking, not calibration or a fixed clinical operating point.

2.6 Reproducibility and computational environment

Both entries implement the official train_model, load_model, and run_model interfaces. Session identifiers are normalized, malformed records are isolated without

terminating extraction, missing inputs are counted, and the saved feature dimension is checked at inference. The entries are network-independent: model artifacts are packaged locally, and TabPFN telemetry and online checkpoint downloads are disabled.

The environment uses Python 3.10.1 with NumPy 2.0.2, pandas 2.2.2, SciPy 1.13.1, scikit-learn 1.6.0, edfio 0.4.10, XGBoost 2.1.1, and TabPFN 7.1.0. Development was performed on a MacBook Pro M4 Max with 36 GB memory. XGBoost uses all CPU cores; the TabPFN code selects CUDA when available and otherwise runs on CPU.

3. Results

Small-set trained XGBoost achieved the best blinded age-conditioned AUROC, 0.765 (ranked 23rd out of 512 successful submissions), compared with 0.716 for large-set XGBoost and 0.645 for small-set TabPFN V2. The complete XGBoost results reveal a metric-dependent reversal. Relative to the small-set model, the large-set model increased conventional AUROC from 0.851 to 0.855, AUPRC from 0.397 to 0.404, and accuracy from 0.913 to 0.931. However, it reduced age-conditioned AUROC by 0.049, age-weighted AUROC by 0.092, reward by 0.126, and F-measure by 0.015. Thus, the large-set model was slightly better at overall case-control discrimination but worse on the age-sensitive primary objective and the submitted binary decisions.

Table 2. Blinded XGBoost results by training-set size.

Metric	Large set	Small set
Reward	0.137	0.263
Age-conditioned AUROC (Challenge Score)	**0.716**	**0.765 (23rd/512)**
Age-weighted AUROC	0.717	0.809
Conventional AUROC	0.855	0.851
AUPRC	0.404	0.397
Accuracy	0.931	0.913
F-measure	0.395	0.410

3.1 Comparative result

The small-set XGBoost pipeline exceeded TabPFN V2 by 0.120 absolute points on the primary metric; large-set XGBoost exceeded it by 0.071. More importantly, the XGBoost size comparison shows that more labeled records did not monotonically improve the target metric. The result favors the small-set submission for the Challenge while leaving open whether the large-set model captures complementary ranking information useful for an ensemble.

4. Discussion

The large-versus-small comparison is more informative than the primary score alone. The large model's higher conventional AUROC, AUPRC, and accuracy indicate that it did not simply learn a uniformly worse classifier. Instead, its ranking changed in a way penalized by age-conditioned evaluation. Because age is itself a predictor, a model can improve ordinary AUROC by separating older positive patients from younger negative patients without improving discrimination between patients of similar age. Additional large-set records may also reinforce site, montage, missingness, or cohort-composition signals that do not transfer to I0004. The higher accuracy but lower F-measure and reward is consistent with a threshold that favored the majority class under the hidden prevalence. These mechanisms remain hypotheses because only one blinded cohort was observed and no confidence intervals or patient-level predictions were available.

XGBoost's advantage over TabPFN is also not an isolated classifier comparison. XGBoost used 156 amplitude-preserving features, whereas TabPFN used 243 normalized features. Plausible contributors include boosted trees' direct fit to thresholded clinical interactions, their ability to ignore weak variables, and the lower effective dimension. TabPFN's synthetic prior may fit correlated PSG summaries poorly under institutional shift; z-scoring made several amplitude summaries nearly fixed; and its human-annotation block was present in training but zero-filled in hidden cohorts where human annotations were unavailable [1]. XGBoost also compared three regularized configurations, whereas TabPFN used one fixed 32-estimator configuration.

For XGBoost, the large set should be treated as a distribution to control rather than an automatic upgrade. Hyperparameters should be retuned separately with nested leave-one-site-out validation using the official age-conditioned scorer. Age-stratified sampling, age-matched positive-negative weighting, an age-excluded residual model, and explicit balancing across sites and channel-availability patterns can test and reduce reliance on confounding structure. Ensembles across balanced large-set subsamples may retain the statistical benefit of 6,600 records while matching the small set's inductive bias. For TabPFN, human features should be removed or masked, raw-amplitude descriptors retained beside normalized features, preprocessing fitted within each fold, and feature subsets and ensemble settings selected by the official metric. Feature-matched XGBoost/TabPFN ablations remain essential.

Late fusion can include all three streams: small-set trained XGBoost, large-set XGBoost, and TabPFN. Rank-transforming site-aware out-of-fold probabilities and selecting weights by age-conditioned AUROC avoids assuming matched calibration; the small-set XGBoost model should receive the dominant initial weight. Fusion is useful only if errors differ, so score correlation and performance on age-matched disagreement pairs should be reported. Thresholds for reward and F-measure should be optimized separately from ranking. A complementary

future stream is SleepFM, a multimodal sleep foundation model pretrained on more than 585,000 hours from approximately 65,000 participants using EEG/EOG, ECG, EMG, and respiratory signals [14]. Frozen SleepFM embeddings could recover temporal morphology discarded by global summaries and be combined with XGBoost or TabPFN predictions, although its 128-Hz input and cross-site transfer would require separate validation.

5. Conclusion

We developed two open-source algorithmic approaches for predicting future cognitive impairment from heterogeneous clinical PSG. On the blinded evaluation, the 156-feature XGBoost system trained using the small dataset achieved an age-conditioned AUROC of 0.765, outperforming the 243-feature TabPFN V2 system at 0.645 and the large-set trained XGBoost model at 0.716. The large-set trained model nevertheless produced slightly higher conventional AUROC, showing that sample size changed which relationships were learned rather than uniformly reducing discrimination. These results support carefully regularized boosted trees while cautioning that more records, more features, or foundation-model complexity need not improve an age-conditioned cross-site objective. Metric-aligned sampling, controlled ablations, deployment-matched missingness, and rank fusion offer practical paths forward; SleepFM embeddings may add information from raw temporal physiology that neither engineered table retains.

Code Availability: Source code for the XGBoost and TabPFN V2 entries is available at https://github.com/emoryeli/insomniac111.git under the BSD 3-Clause license.

Acknowledgments

I thank the previous PhysioNet Challenge participants for many helpful ideas in their papers and shared codes. I also want to thank the George B. Moody PhysioNet Challenge organizers, the Human Sleep Project contributors, and the developers of CAISR, XGBoost, and TabPFN for making the data, annotations, and source codes available.

References

[1] Reyna MA, et al. Screening for Cognitive Impairment During Sleep Studies: The George B. Moody PhysioNet Challenge 2026. https://moody-challenge.physionet.org/2026/ (accessed 25 Aug 2026).

[2] Miller MA. The role of sleep and sleep disorders in the development, diagnosis, and management of neurocognitive disorders. Front Neurol. 2015;6:224. doi:10.3389/fneur.2015.00224.

[3] Li Q, et al. The Human Sleep Project: a multi-center clinical polysomnography dataset across the human lifespan. Sleep. 2026;zsag215. doi:10.1093/sleep/zsag215.

[4] 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.

[5] Goldberger AL, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000;101:e215-e220.

[6] Vaswani A, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30.

[7] Hollmann N, Muller S, Eggensperger K, Hutter F. TabPFN: a transformer that solves small tabular classification problems in a second. ICLR. 2023. arXiv:2207.01848.

[8] Hollmann N, et al. Accurate predictions on small data with a tabular foundation model. Nature. 2025;637:319-326. doi:10.1038/s41586-024-08328-6.

[9] Grinsztajn L, et al. TabPFN-2.5: advancing the state of the art in tabular foundation models. arXiv:2511.08667v2. 2026.

[10] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc KDD. 2016:785-794. doi:10.1145/2939672.2939785.

[11] Woo S, et al. ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders. Proc CVPR. 2023:16133-16142. arXiv:2301.00808.

[12] Ramar K, et al. Sleep is essential to health: an American Academy of Sleep Medicine position statement. J Clin Sleep Med. 2021;17:2115-2119.

[13] Hjorth B. EEG analysis based on time domain properties. Electroencephalogr Clin Neurophysiol. 1970;29:306-310.

[14] Thapa R, Kjaer MR, He B, et al. A multimodal sleep foundation model for disease prediction. Nat Med. 2026;32:752-762. doi:10.1038/s41591-025-04133-4.

Address for correspondence:
Elaine Li
College of Arts and Sciences, Emory University, Atlanta, Georgia, USA
elaine.li@emory.edu