Cross-site generalization of interpretable sleep features for predicting future cognitive impairment

S. A. I. Shouborno[1], Bashima Islam[1]

[1]Department of Electrical and Computer Engineering, Riccio College of Engineering, University of Massachusetts Amherst, Amherst, MA, USA

Abstract

As part of Screening for Cognitive Impairment During Sleep Studies: The George B. Moody PhysioNet Challenge 2026, our team, bashlab_wpi, predicted future diagnoses of cognitive impairment from overnight polysomnography (PSG) with interpretable features, concentrating the modelling effort on transfer to an unseen site.

Each recording is reduced to 151 named features, none of them learned: 86 classical PSG summaries, 63 that summarise the distribution of band power in 30 s epochs rather than its whole-night mean, and 2 recording-date features. These are aligned to the scored site by correlation alignment (CORAL), which uses the unlabelled target recordings, then scored by a rank blend of TabFM, a tabular foundation model applied in context, weighted 0.7*, and a tree weighted* 0.3*.*

The largest single increment, 0.092, comes from recording date rather than from sleep. A negative label needs six years of follow-up and a positive as little as one, and all follow-up ends when the data are assembled, so after 2020 the set holds 38 positives and no negatives. With the date already in the model the classical features still add +0.036 and the pooling features +0.003. Leave-one-site-out cross-validation, not comparable to the official scores, spans 0.692–0.860. On the hidden validation set the entry received age-conditioned AUROC 0.751, ranked 29/514.

1. Introduction

Cognitive impairment is often recognised years after the underlying process begins, while polysomnography (PSG) is already recorded at scale for other indications [1], so a screening signal read from an existing study would need no new acquisition.

We participated in the 2026 George B. Moody PhysioNet Challenge, which invited teams to predict future diagnoses of cognitive impairment from PSG [1, 2], on recordings from the Human Sleep Project [3]. Scoring is age-conditioned, and the scored recordings come from one hospital absent from training. What limits a model here is therefore not how well it separates cases within a familiar site, but how much of that separation it keeps at a site it has never seen.

This paper contributes an interpretable feature set for the task, a label-free alignment step for the unseen scored site, and a measurement showing that recording date, rather than sleep, supplies the largest single signal in these data.

2. Methods

Each recording is reduced to 151 named features. Those features are transformed so their covariance matches the site being scored, and the two models fitted on them are combined by averaging the ranks they assign, as Section 2.4 sets out.

2.1. Data

The public training set holds 6,600 recordings from three sites (I0002, I0006 and S0001) with 498 positives, a prevalence of 7.55%. Every labelled recording entered training, with no exclusion criteria, no relabelling and no external data. Records were processed independently and no patient spans a split.

Body mass index is missing for 76.7% of records and the missingness is site-dependent, so we dropped the column rather than impute it.

2.2. Features

Every feature is a named physiological quantity or a property of the recording, so any effect the model finds can be traced to something interpretable. Three blocks supply 151 of them.

Classical PSG summaries (86 features). Staging, arousal and respiratory-event summaries from CAISR [4]; the 5×5 stage-transition matrix; per-stage bout statistics and spectral summaries; heart rate variability via NeuroKit2 [5]; slow-wave and spindle events via YASA [6]; oxygen saturation and demographics.

Temporal pooling (63 features). Classical features reduce a night to a mean, which cannot separate an unstable night from a uniformly poor one. Following the winning entry of the 2023 I-CARE Challenge [7], the closest analogous task, which aggregated per-segment features by quantiles, this block instead summarises the distribution of per-epoch band power across the night. Relative power in five bands (delta 0.5–4, theta 4–8, alpha 8–12, sigma 12–16, beta 16–30 Hz) is computed by Welch's method on each 30 s epoch of one electroencephalogram derivation, taken from a fixed priority order, over epochs CAISR scored as sleep. Each band contributes twelve statistics: the 10th, 25th, 50th, 75th, 88th and 95th percentiles, the interquartile and 10–90 ranges, standard deviation over median, last third of the night minus first, the largest share of any one-hour window spent above the 88th percentile, and upward crossings of that percentile per hour. Three ratios of the 88th percentiles, delta over alpha, theta over alpha and delta over beta, complete the block: $5 \times 12 + 3 = 63$.

Recording date (2 features). A positive label requires a diagnosis one to six years after the recording and a negative requires six or more years with no diagnosis, which makes the date informative for the reason Section 4.2 sets out, so it enters as recording year t_i in fractional years and as the follow-up the record could have accrued had its site's last recording been followed for six years, $d_i = \max_{j \in s(i)} t_j + 6 - t_i$, where $s(i)$ is the site of record i and the maximum runs over all recordings at that site. The offset of 6 years matches the follow-up interval in the label definition. Both are read from the demographics file, not the signal.

Missing values are filled with training-set medians. The tree standardises the features when it is fitted before alignment; the models refitted after it use no scaler, so Eq. (1) and its ridge act on the features' native scales. The 151 features were chosen in advance from a pool of 1,015 candidates, which also held magnitude-squared coherence between derivations and a block derived from the recording's own file structure. Both were dropped: neither improved the cross-validated mean, and the second is a property of how a study was stored rather than of the patient. The choice was made before the runs reported here, and no selector runs during training or inference.

2.3. Alignment to the scored site

The scored site is absent from training, but its recordings are available unlabelled before any prediction, which is what correlation alignment [8] exploits. With C_s, C_t the feature covariance matrices and μ_s, μ_t the feature means of the training and target sets, the training matrix X_s is whitened and recoloured to

$$\tilde{X}_\mathrm{s} = \big(X_\mathrm{s} - \mu_\mathrm{s}\big) C_\mathrm{s}^{-1/2} C_\mathrm{t}^{1/2} + \mu_\mathrm{t}, \qquad (1)$$

with a ridge of 10^{-4} added to both covariances. Second-order structure is matched without any label, and both models are then refitted on $\tilde{X}_\mathrm{s}$ at inference. Below 200 target records a 151×151 covariance is too poorly conditioned to help, so alignment is skipped and both unaligned models are used. The target matrix is imputed before its covariance is taken; otherwise the covariance is not finite and Eq. (1) returns the training features unchanged.

2.4. Models and blending

A gradient boosted decision tree [9] and TabFM [10], a tabular foundation model, are fitted on the same aligned features. TabFM is applied *in context*, carrying the training set through every forward pass rather than distilling it into weights, with the site identifier as an extra covariate and a reserved code for the unseen site met at scoring time.

The two are combined by rank, because the metric reads only order while the models are calibrated differently. For n records with probabilities p^F from TabFM and p^G from the tree, the blended score is $s_i = w\,\mathrm{rank}(\mathrm{p^F})_\mathrm{i}/\mathrm{n} + (1 - \mathrm{w})\,\mathrm{rank}(\mathrm{p^G})_\mathrm{i}/\mathrm{n}$ with $w = 0.7$.

The Challenge takes two outputs per record, and they are built differently because the two metrics reward different things. The ranking metric reads only order, so the probability is the blend above. The reward is paid on the binary decision, which needs a probability that can be compared against how common the condition is at a given age: the decision therefore thresholds the tree's posterior, shifted from the training prevalence to an assumed 0.10, against the observed prevalence within ± 2 years of that record's age.

Inference cost is nearly independent of the number of records, so all are scored on the first call and cached.

2.5. Parameters

Table 1 lists every setting. Three were swept by leave-one-site-out cross-validation and the rest were fixed once, without a search. The blend-weight sweep did not separate $w = 0.7$ from TabFM alone by more than 0.005, so the two are tied on our evidence and 0.7 is a choice within that tie. The 200-record floor on alignment was set in advance on conditioning grounds and is never reached in cross-validation, where the smallest target fold has 319 records.

2.6. Model selection and evaluation

Selection ran by leave-one-site-out cross-validation on the public training set: each site is held out in turn, both models are fitted on the remaining two and aligned to the held-out site, and age-conditioned AUROC is computed

Setting	Value	Swept
Blend weight w on TabFM	0.7	yes
Tree estimators, learning rate	300, 0.03	yes
Tree leaves, min. per leaf	15, 50	no
Tree subsample, L_2 penalty	0.5, 1.0	no
TabFM members, max. features	32, 500	no
TabFM SVD features	square root	no
CORAL ridge, min. records	10^{-4}, 200	no
Imputation, random seed	median, 42	no

Table 1. All parameters of the submitted entry. 'Swept' marks the three chosen by leave-one-site-out cross-validation; the rest were fixed without a search. 'square root' sets the singular value decomposition feature count to the square root of the feature count. CORAL: correlation alignment; TabFM: tabular foundation model.

I0002	I0006	S0001	Validation	Test	Ranking
0.860	0.795	0.692	0.751	—	29/514

Table 2. Challenge scores for our selected entry (team bashlab_wpi), with the ranking on the hidden validation set, to be replaced by the test-set figures once final scores are released. The per-site columns are per-site cross validation on the public training set and are not comparable to the official columns; the test column is blank pending one-time scoring on the hidden test set. All entries are age-conditioned AUROC (n.u.) at a matching gap of 2 years.

there at a matching gap of 2 years. A single quoted figure is the unweighted mean over the three folds and is not comparable to the official scores. Repeats within one job are bit-identical, but the same configuration run on different devices differs by 0.008 on a fold against 0.004 on the mean, since in-context inference is not bitwise reproducible across devices. We therefore select on the mean, treat differences below 0.005 as indistinguishable, and report each fold separately, since a model scored on one unseen site is exposed to whichever fold that site resembles.

The extraction, training and evaluation code is public at `github.com/shouborno/physionet-challenge-2026`.

3. Results

On the hidden validation set the entry received age-conditioned AUROC 0.751 at rank 29/514 (Table 2), with plain AUROC 0.887, area under the precision-recall curve 0.489, accuracy 0.664 and reward 0.118, the Challenge's cost-weighted score on the binary decisions.

The metric scores age-matched record pairs, and the folds differ in pair count by two orders of magnitude, 1,886 for I0002 against 182,330 for S0001, with bootstrap 95% intervals [0.78–0.93] and [0.65–0.73].

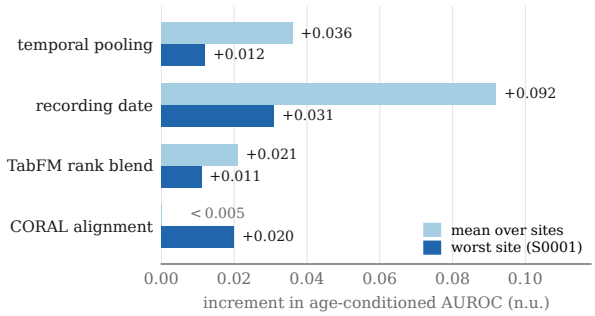

Figure 1. Increment in age-conditioned AUROC (n.u.) from each block, added in order, to the mean over the three leave-one-site-out folds and to the worst fold. A value too small to separate from zero, given how far a rerun moves it, is labelled as a bound instead of a number. Blocks above the TabFM row use the tree alone. AUROC: area under the receiver operating characteristic curve; CORAL: correlation alignment; TabFM: tabular foundation model.

3.1. Contribution of each block

Figure 1 gives each block's increment to the mean over the three folds and to the score on the worst fold. Recording date supplies +0.092 to the mean and alignment +0.020 to the worst fold.

3.2. Alignment

We measured alignment twice, in a paired run against an unaligned arm and again inside the ladder. Both raise the hardest site, S0001, by +0.042 and $+0.020$, and both lower I0002, the site with the widest confidence interval, by 0.041 and 0.030. The mean is unchanged within our resolution, +0.003 and $+0.000$, both below 0.005.

4. Discussion and Conclusions

4.1. Transfer to the scored site

Named interpretable features transferred to a hospital absent from training. The validation score falls inside the interval the three folds span, above the pair-count weighted mean 0.701 and below the unweighted mean 0.782. An interval that wide makes containment a sanity check rather than evidence the folds are calibrated, but it does rule out the failure we were guarding against: strength on the two large training sites and collapse on an unseen one.

Alignment raises the hardest site in both measurements at no label cost, which is why we keep it, though the mean does not move beyond our resolution. Which site gains and which loses follows pair count, but with only three sites

it follows record count, collection window and prevalence just as well, so we cannot say which of them is responsible.

The model is usable at a screening budget on its weakest fold: referring the top 10% by score on S0001 finds 41% of the future diagnoses there, so a clinician reads 3.8 studies for every case found against 15.4 at that site's base rate. Recording year alone reaches 34% at the same budget, so the physiology accounts for the remainder.

4.2. What recording date measures

A positive label requires a diagnosis one to six years after the recording and a negative requires six or more years with no diagnosis. All of that follow-up has to be complete when the dataset is assembled, and here it ends around 2023, so the classes cannot occupy the same span of recording dates. The rule binds exactly, with no negative below six years, and the effect is stark: after 2020 the set holds 38 positives and no negatives at all, leaving negative recordings older by a median of 3.3 years. The same labelling code generates the hidden sites' labels, so the same bias should be present there, and the entry therefore keeps the feature.

How far the date carries depends on a site's collection window, and it already varies among the training sites: date alone separates cases from controls at 0.811 in I0002, whose recordings span 6.7 years, but only 0.660 in S0001, which spans 14.7 years. Figure 1 adds the physiology before the date; adding it after the date instead separates the two. From date alone at 0.723, the 86 classical features add +0.036 to the mean and +0.024 to the worst fold. The 63 pooling features then add +0.003 and +0.004, inside our resolution, where the same block is worth +0.036 in Figure 1 because there it is measured before the date goes in. Temporal pooling was largely standing in for the date.

4.3. Limitations

Alignment was measured across the three training sites and on a synthetic widening of the gap between them, but how far the scored site really sits from training is something we could not measure. Our accuracy 0.664 and F-measure 0.203 are low beside the ranking metric, because the decision threshold adapts to age but not to site and assumes a fixed 0.10 target prevalence.

4.4. Conclusions

Interpretable features with label-free alignment reached 0.751 on an unseen hospital, inside the range that holding whole sites out had predicted. Recording date is the largest single signal in these data, informative through the label definition rather than through sleep, and it accounts for most of the screening enrichment above. Because a physiological block's apparent contribution depends on whether the date is already in the model, we report both orderings.

Acknowledgments

No external funding. We declare no conflicts of interest.

References

[1] Reyna MA, Weigle A, Li Q, Koscova Z, Sun HS, Wen S, et al. Screening for Cognitive Impairment During Sleep Studies: The George B. Moody PhysioNet Challenge 2026. In Computing in Cardiology 2026, volume 53. 2026; 1–3.

[2] Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 2000; 101(23):e215–e220.

[3] Li Q, Wen S, Sun H, Ganglberger W, Tripathi A, Turley N, et al. The Human Sleep Project: A multi-center clinical polysomnography dataset across the human lifespan. Sleep 2026;zsag215. ISSN 0161-8105.

[4] 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.

[5] Makowski D, Pham T, Lau ZJ, Brammer JC, Lespinasse F, Pham H, et al. NeuroKit2: A Python toolbox for neurophysiological signal processing. Behavior Research Methods 2021;53(4):1689–1696.

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

[7] Zabihi M, Chaman Zar A, Grover P, Rosenthal ES. HyperEnsemble learning from multimodal biosignals to robustly predict functional outcome after cardiac arrest. In Computing in Cardiology, volume 50. 2023; 1–4.

[8] Sun B, Feng J, Saenko K. Return of frustratingly easy domain adaptation. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, volume 30. 2016; 2058–2065.

[9] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, volume 30. 2017; 3146–3154.

[10] Kong W, Das A. TabFM: A zero-shot foundation model for tabular data. Google Research, 6 2026. Model and code at `github.com/google-research/tabfm`.

Address for correspondence:

S. A. I. Shouborno
University of Massachusetts Amherst, MA 01003, USA
sasifimran@umass.edu