Age-Conditioned Evaluation of Handcrafted Polysomnography Models for Cognitive-Impairment Prediction

Satwik Kapavarapu[1,*], Nathan Hoang[2,*], Andrei Malagusan[2,*]

[1] University of Washington, Seattle, WA, USA
[2] Georgia Institute of Technology, Atlanta, GA, USA
[*] These authors contributed equally to this work.

Abstract

Sleep disturbances have been associated with cognitive decline, but generalizable prediction from routine polysomnography remains difficult. As team gdub, we developed a CPU-oriented system that represented each study with 830 handcrafted demographic, physiological, and annotation-derived features. Training-cohort preprocessing removed configured BMI and coverage/missingness fields, excluded the 30% of candidate features with the largest cross-site Kolmogorov–Smirnov statistics, and applied median imputation. ExtraTrees, CatBoost, and LightGBM probabilities were averaged and calibrated with out-of-fold isotonic regression; the submitted system also included an age-binned LambdaRank score and an age-local prevalence rule for binary predictions. In strict public three-site leave-one-site-out (LOSO) evaluation of the classifier path with LambdaRank disabled, the submitted KS setting achieved a macro age-conditioned AUROC of 0.654. Official hidden validation of submission 2163 produced an age-conditioned AUROC of 0.486 despite an ordinary AUROC of 0.746. This discrepancy indicates that all-pair discrimination did not persist among age-comparable participants and identifies age-conditioned cross-source generalization as the principal limitation of the submitted approach.

1. Introduction

Cognitive impairment, including mild cognitive impairment and dementia, is commonly assessed through cognitive testing or neuroimaging, approaches that can be difficult to deploy at population scale. Disrupted slow-wave sleep and sleep-disordered breathing or nocturnal hypoxemia have been associated with cognitive decline [1, 2], motivating investigation of routinely collected polysomnography (PSG) as a potential risk marker.

The George B. Moody PhysioNet Challenge 2026 asked participants to predict future cognitive-impairment diagnoses from PSG recordings [3]. Its primary age-conditioned AUROC restricts positive–negative comparisons to participants of similar age. This endpoint can expose failures hidden by ordinary AUROC when predictions rely on age-associated or source-associated structure. Heterogeneous channel configurations and clinical sources further make cross-site generalization central to model development.

This work makes three contributions. First, it describes the exact self-contained 830-feature package evaluated as gdub submission 2163. Second, it compares conventional and age-conditioned discrimination on official hidden validation data. Third, it examines public-site sensitivity to cross-site KS filtering and uses the resulting development-to-hidden gap to characterize limitations of source-robust model selection.

2. Methods

2.1. Data and Preprocessing

Organizer data included overnight PSG, demographics, CAISR sleep annotations [4], human annotations, and binary labels [3]. Standardized channel aliases supported available EEG, EOG, chin/leg EMG, ECG or heart rate, airflow, thoracic/abdominal effort, and oxygen saturation; missing modalities did not exclude studies.

Human-annotation extraction was disabled, leaving 18 reserved positions zero-filled. Configured coverage/missingness fields were removed, and deployment preprocessing was fitted only on the training cohort and applied unchanged at hidden inference.

2.2. Handcrafted Feature Extraction

Table 1 summarizes the 830-position schema. The 800-position base vector comprised demographic fields, multimodal physiological summaries, and CAISR stage, architecture, event, arousal, limb-movement, and sequence measures.

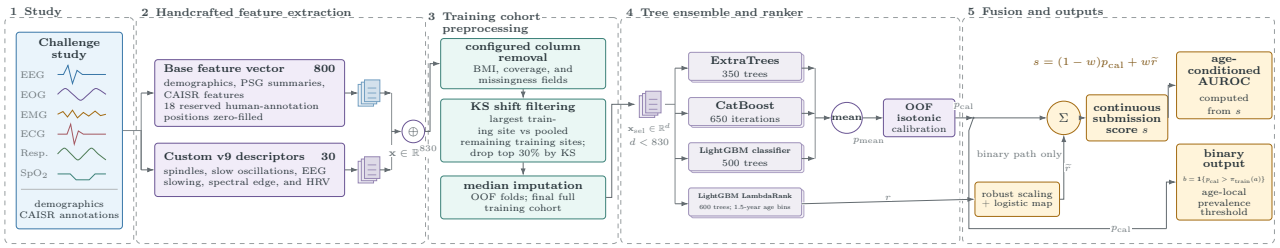

Figure 1: Pipeline for gdub submission 2163. The 830-feature representation undergoes training-cohort filtering and median imputation, tree-ensemble calibration, optional age-binned LambdaRank blending, and separate continuous and binary output paths.

Table 1: Exact ordered feature schema before configured column removal. Human-annotation positions were retained but zero-filled.

Block	Dim.	Contents
Demographic	16	Age, sex, BMI, race, ethnicity, and missingness encodings
Physiological	455	Multimodal PSG, stage EEG, quality, coverage, and relative summaries
Algorithmic annotations	311	Architecture, events, and annotation sequence summaries
Reserved human annotations	18	Fixed schema positions; zero-filled
v9 descriptors	30	Spindles, slow oscillations, slowing, spectral edge, HR/HRV, and coverage

The 30 v9 positions described N2/N3 spindles, N3 slow oscillations and coupling, EEG slowing, spectral-edge frequency, heart rate/variability, and coverage. Thus,

$$\mathbf{x} = [\mathbf{x}_{\text{base}} \| \mathbf{x}_{\text{v9}}] \in \mathbb{R}^{800+30} = \mathbb{R}^{830}.$$

Here, $\mathbf{x}_{\text{base}}$ and $\mathbf{x}_{\text{v9}}$ are the 800- and 30-position blocks; 16 + 455 + 311 + 18 + 30 = 830.

2.3. Training-Cohort Site-Shift Filtering

A training-cohort keep-mask removed BMI, its missingness indicator, and configured coverage/missingness fields. Two-sample Kolmogorov–Smirnov statistics compared finite observations from the largest site with pooled observations from all others; approximately the highest 30% were excluded.

The mask was applied unchanged to hidden records and did not establish invariance. Its reuse within the five-fold out-of-fold (OOF) construction loop means those OOF quantities are not unbiased generalization estimates; hidden validation data were not used.

2.4. Tree Ensemble and Probability Calibration

With seed 2026, the safe profile fitted ExtraTrees [5] (350 trees, square-root features, leaf size 2, balanced-subsample weighting), CatBoost [6] (650 iterations, rate 0.03, depth 6, L2 5.0, negative-to-positive class weighting), and LightGBM [7] (500 trees, rate 0.025, 31 leaves, child size 20, row/column subsampling 0.80, L2 5.0, positive-class ratio weighting).

Five patient-grouped, label-stratified folds used fold-fitted median imputers to generate OOF predictions. Isotonic regression [8] calibrated their mean; final models used all training rows:

$$q_i = \frac{1}{3}\left(p_i^{\text{ET}} + p_i^{\text{CB}} + p_i^{\text{LGBM}}\right), \qquad p_i^{\text{cal}} = \text{Iso}(q_i),$$

The three p_i terms are classifier probabilities, q_i is their mean, and Iso is the OOF-fitted map; calibration was not assumed to improve ranking AUROC.

2.5. Age-Aware Ranking and Binary Output

The LightGBM LambdaRank path [9] used 1.5-year age bins and 600 estimators. Robustly scaled rank outputs were blended as

$$z_i = (1 - w)p_i^{\text{cal}} + wr_i.$$

Here, r_i is the rank score and z_i the continuous score. Weight w was selected from the submitted grid using training-cohort OOF age-conditioned AUROC; no fixed value or LambdaRank-enabled public LOSO score is reported.

Binary output compared p_i^{cal} with training prevalence within ±2 years of age a_i:

$$\widehat{y}_i = \mathbf{1}\left[p_i^{\text{cal}} > \widehat{\pi}_{\text{train}}(a_i)\right].$$

Here, $\widehat{y}_i$ is the binary prediction and $\mathbf{1}[\cdot]$ is the indicator. This threshold affected binary metrics, not continuous ordering.

2.6. Evaluation Protocols and Metric

Public three-site LOSO recomputed filtering, imputation, calibration, and classifier fitting using only training sites; held-out sites were evaluated with LambdaRank disabled. Organizers evaluated the complete submission on hidden data [3].

For continuous score z_i, label y_i, and age a_i,

$$\mathcal{P} = \{(i, j) : y_i = 1,\ y_j = 0,\ |a_i - a_j| \le 2\},$$

$$\mathrm{AUROC}_{\mathrm{age}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[\mathbf{1}(z_i > z_j) + \tfrac{1}{2}\mathbf{1}(z_i = z_j)\right].$$

Here, $\mathbf{1}(\cdot)$ is the indicator, with half credit for ties. Unlike ordinary AUROC, the metric evaluates ordering only within age-comparable pairs.

3. Results

3.1. Public LOSO Evaluation

Table 2 separates fold-isolated public evaluation from hidden validation. At KS=0.30 with LambdaRank disabled, public LOSO achieved macro age-conditioned AUROC 0.654; site values were 0.665, 0.706, and 0.591 for I0002, I0006, and S0001, with pooled AUROC 0.625. The 0.168 public-to-official difference is descriptive because the score paths differed and the source composition and source-specific official metrics were unavailable.

3.2. KS-Filter Sensitivity

Table 3 reports classifier-only public LOSO sensitivity. Relative to no filtering, submitted KS=0.30 increased macro AUROC from 0.640 to 0.654. KS=0.20 had the highest retrospective macro value (0.662), while KS=0.25 had the highest worst-site value (0.599); the official package used 0.30. S0001 remained weakest at that setting, so marginal screening did not yield uniformly robust performance.

These are development sensitivity results; retrospectively selecting the best row would not be unbiased model selection.

3.3. Official Validation

Table 4 reports the primary external result. Submission 2163 achieved age-conditioned AUROC 0.486 despite ordinary AUROC 0.746, a 0.260 difference. The model retained all-pair discrimination but did not reliably order participants within age-comparable pairs. This is consistent with age- or source-associated structure, although aggregate feedback cannot identify the cause. Reward, accuracy, and F-measure also indicated weak thresholded performance.

Patient-level predictions, confidence intervals, hidden-source identifiers, site metrics, and hidden calibration curves were unavailable.

3.4. Reproducibility and Scope

This paper describes the CPU-oriented safe profile evaluated as gdub submission 2163 at commit `a7f966e2ddfa3c8ecbe9a05cb9cdfa9c260d7889`. It reconstructed features from raw Challenge-format inputs without a local cache or external pretrained checkpoint. The paper concerns submission 2163 only, not later development systems. Official results are aggregate point estimates; this retrospective evaluation does not establish clinical validity.

4. Discussion and Conclusion

4.1. Interpretation

The central observation is the 0.260 gap between ordinary AUROC (0.746), which ranks all positive–negative pairs, and age-conditioned AUROC (0.486), which restricts pairs by age. Thus, conventional discrimination did not persist on the target endpoint, and accuracy was also misleading under class imbalance.

Classifier-only public LOSO exceeded hidden official performance by approximately 0.168. This is consistent with source shift or model-selection instability, but protocols differed because public LOSO disabled LambdaRank and hidden predictions were unavailable. KS filtering modestly improved public macro performance, yet S0001 remained difficult, showing that marginal screening was insufficient.

Model selection should use the exact endpoint with fold-local preprocessing and rank-weight selection, and report per-source and worst-source results. Hidden evaluation remains essential; shift-focused methods do not consistently transfer under realistic model selection [10].

4.2. Limitations

Official feedback contained only aggregate estimates, without patient predictions, confidence intervals, or site metrics. Public LOSO covered three sites and disabled LambdaRank. The artifact's keep-mask was not fold-local during OOF construction; marginal KS screening could retain multivariate site signatures or remove biological variation. Matched ablations did not isolate the v9 block, learners, calibration, or LambdaRank, whose blend used training-cohort OOF predictions. Class imbalance limited accuracy, and no clinical-utility claim is supported.

Table 2: Public LOSO recomputed the full classifier pipeline within each training-site split and disabled LambdaRank. The official row is the organizer-reported result for the complete package; hidden source metrics were unavailable. "Overall" denotes the macro public or aggregate official age-conditioned AUROC.

Evaluation	Model path	Overall	I0002	I0006	S0001	Pooled
Public three-site LOSO	Classifier, KS=0.30, LambdaRank off	0.654	0.665	0.706	0.591	0.625
Official hidden validation	Full submission 2163	0.486	–	–	–	–

Table 3: Public classifier-only LOSO sensitivity with LambdaRank disabled. BMI and coverage were removed except in the final row; the dagger marks the submitted KS setting.

Configuration	Macro	Worst	Pooled
No KS; drop BMI/cov.	0.640	0.553	0.591
KS=0.20; drop BMI/cov.	0.662	0.590	0.619
KS=0.25; drop BMI/cov.	0.661	0.599	0.628
KS=0.30; drop BMI/cov.†	0.654	0.591	0.625
KS=0.30; retain coverage	0.651	0.585	0.615

Table 4: Organizer-reported aggregate metrics for gdub submission 2163 on hidden large-dataset validation.

Metric	Value	Metric	Value
Reward	−0.186	AUPRC	0.201
Age-conditioned AUROC	0.486	Accuracy	0.424
Age-weighted AUROC	0.444	F-measure	0.077
Ordinary AUROC	0.746		

4.3. Future Work and Conclusion

Future work should fully nest preprocessing and rank-weight selection; isolate the v9 block, learners, calibration, and LambdaRank; select for source-balanced or worst-source performance; and evaluate new acquisition protocols with grouped bootstrap intervals when predictions are available.

In conclusion, the classifier path achieved public LOSO macro age-conditioned AUROC 0.654 with LambdaRank disabled, whereas the complete submission achieved 0.486 on hidden validation despite ordinary AUROC 0.746. Conventional discrimination was therefore insufficient for robust within-age ranking, emphasizing fully nested, source-aware model selection.

Acknowledgments

The authors thank the Challenge organizers and data contributors.

Conflict of interest: The authors declare no conflicts of interest.

Code and data availability: The code corresponding to submission 2163 is available to the Challenge organizers at commit `a7f966e2ddfa3c8ecbe9a05cb9cdfa9c260d7889`. Challenge data access is governed by the organizers and the data are not redistributed.

Correspondence: Raghavendra Satwik Kapavarapu, rkapav@uw.edu.

References

[1] B. A. Mander, V. Rao, B. Lu, *et al.*, "Prefrontal atrophy, disrupted NREM slow waves and impaired hippocampal-dependent memory in aging," *Nature Neuroscience*, vol. 16, pp. 357–364, 2013, doi: 10.1038/nn.3324.

[2] K. Yaffe, A. M. Laffan, S. L. Harrison, *et al.*, "Sleep-disordered breathing, hypoxia, and risk of mild cognitive impairment and dementia in older women," *JAMA*, vol. 306, no. 6, pp. 613–619, 2011, doi: 10.1001/jama.2011.1115.

[3] George B. Moody PhysioNet Challenge, "George B. Moody PhysioNet Challenge 2026," PhysioNet, 2026. [Online]. Available: https://moody-challenge.physionet.org/2026. [Accessed: Aug. 30, 2026].

[4] S. Nasiri *et al.*, "CAISR: achieving human-level performance in automated sleep analysis across all clinical sleep metrics," *Sleep*, vol. 48, no. 8, Art. no. zsaf134, 2025, doi: 10.1093/sleep/zsaf134.

[5] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[6] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[7] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 3149–3157, 2017.

[8] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Machine Learning*, pp. 625–632, 2005, doi: 10.1145/1102351.1102430.

[9] C. J. C. Burges, "From RankNet to LambdaRank to LambdaMART: An overview," Microsoft Research, Tech. Rep. MSR-TR-2010-82, 2010.

[10] I. Gulrajani and D. Lopez-Paz, "In Search of Lost Domain Generalization," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://arxiv.org/abs/2007.01434.