Site-Aware Stacking for Cross-Center Cognitive Risk Modeling From Overnight Polysomnography

Yuxuan Xu

College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

Abstract

Team GT NeuroSignals Lab submitted an ensemble to rank cognitive-impairment risk one to six years after overnight polysomnography within the organizer's time limit. Four feature groups combined demographics with signal summaries, interrupted-sleep measures, stage transitions, and clinical sleep measures. For each group, models trained on all but one data-collection site predicted records from the excluded site. If time allowed, logistic regression combined their probabilities with variables derived from the database record-creation time; otherwise, the program returned the last completed model. Organizer-side training used a deterministic 1,200-record cap. In validation, entry 2029 returned age-conditioned, age-weighted, and overall areas under the receiver operating characteristic curve of 0.748, 0.813, and 0.790, respectively, and a precision–recall area of 0.366. The prevalence-based reward was 0.022, accuracy 0.382, and F-measure 0.167. The organizers did not return the fitted model or training log, so the scored model is unknown. Final assessment requires a single official test-set evaluation of the selected, unchanged procedure.

1. Introduction

Sleep physiology may provide information beyond conventional sleep-disorder indices. A longitudinal study associated quantitative sleep electroencephalographic (EEG) patterns with later cognitive decline [1]. Cross-site prediction is difficult because polysomnographic channel sets, acquisition systems, annotation availability, and outcome prevalence differ across clinical sites.

The 2026 George B. Moody evaluation asks algorithms to rank patients by the risk of receiving a cognitive-impairment diagnosis one to six years after an overnight polysomnogram (PSG) [2]. The primary measure compares a positive and a negative patient only when their ages differ by no more than two years. We designed four groups of manually defined sleep measures that can be computed on a central processing unit. Each site's records were predicted by models trained only on the other sites, and logistic regression combined their probabilities using stacked generalization [3]. The program saved a model after each group so it could return an earlier model if later fitting exceeded the time limit. We report the submitted procedure and all validation metrics returned by the organizers, while separating measured performance from unresolved implementation details.

2. Methods

2.1. Data, outcome, and sampling

The official large training set contains 6,600 recordings from three sites (S0001, I0002, and I0006), drawn from the Human Sleep Project [4]. Available inputs include raw polysomnography, demographics, automated Complete Artificial Intelligence Sleep Report (CAISR) annotations, and training-only outcomes. CAISR provides sleep stages and arousal, respiratory-event, and limb-movement annotations [5]. A positive outcome requires at least two qualifying diagnoses, separated by at least one week, with the first occurring one to six years after the sleep study. A negative outcome requires no qualifying diagnosis and at least six years of follow-up [2]. Records without either label were excluded from fitting.

To bound repeated feature extraction, the submitted program shuffled record indices with seed 20260619 and selected at most 1,200 records, preserving the same subset and row order for all models. This time-saving sample was not stratified by outcome or site. For each model fold and the final fit, sample weights combined inverse site-frequency and inverse class-frequency factors. They were normalized to mean one, multiplied by a follow-up reliability factor when each relevant class had at least eight finite values, and normalized again. The follow-up multipliers ranged from 0.95 to 1.10 for positive records and 0.90 to 1.10 for negative records; site-specific fits did not use these weights. The follow-up fields entered only this weighting calculation and were not model predictors.

2.2. Complementary feature groups

Each model included age, sex, race, ethnicity, body mass index (BMI), polynomial age and BMI terms, an age–BMI interaction, and explicit missingness indicators. CAISR annotations were used when available; human annotations were used when the corresponding CAISR files were unavailable. The hidden validation and test sets provide algorithmic rather than human annotation files.

The training schedule attempted five feature groups in a fixed order. The first four did not require pretrained models. First, the *signal group* added relative EEG band powers, spectral entropy, non-rapid-eye-movement stage N2 spindle and K-complex density proxies, stage N3 slow-wave proxies, contrasts among N2, N3, rapid-eye-movement (REM), and wake states, and REM-related electrooculographic (EOG) activity. These summaries were computed from selected available channels after median centering and median-absolute-deviation scaling. Second, the *interrupted-sleep group* summarized short stage runs, transition instability, and changes in CAISR stage probabilities. Third, the *sleep-stage-change group* retained a narrower set of transition and probability-change measures. Fourth, the *clinical-sleep group* summarized stage fractions, bout lengths, transition counts, sleep continuity, staging confidence, and the frequency and duration of arousal, respiratory, and limb-movement events.

The fifth feature group requested numerical features from a pretrained brain-age model, but its model files and configuration were absent. On the 1,200-record path, more than 95% of attempts would fail, so this group was omitted. Failure of one group did not stop the remaining fits.

The program first saved a model that assigned every patient the same training-set risk. After each feature group, it saved an equal-weight logistic combination of all completed groups. A timeout stopped later groups; if too little time remained for final stacking, the program returned the last saved model. Because the organizers did not return the fitted model files or log, the ranking metrics can exclude a constant prediction but cannot identify which later model ran.

2.3. Base models and stacking

For the signal group, a three-member histogram-gradient-boosting ensemble (learning rate 0.03, 300 iterations, depth 3, minimum leaf size 10) was paired with elastic-net logistic regression ($C = 0.15$, with equal weights on the L1 and L2 penalties). For each of the annotation-based groups, the candidate set comprised gradient boosting (250 trees, learning rate 0.03, depth 3), 600 extremely randomized trees (minimum leaf size 3), and elastic-net logistic regression ($C = 0.25$, with an L1 penalty fraction of 0.25) [6–8]. The signal-group elastic-net model and the annotation-based models median-imputed continuous features; only the latter added flags marking imputed values. Histogram gradient boosting used its native missing-value handling.

Base-model probabilities used leave-one-site-out cross-validation: each site was predicted by models trained on the others. Before those folds, a screen that did not use outcome labels removed predefined columns that might identify the data source, plus columns with fewer than three observations or negligible variance. Feature availability was therefore determined on the selected 1,200-record matrix before site folds. Each base-model fit excluded the held-out site, but weights within each feature group used pooled predictions for excluded records and their labels. These non-negative weights optimized overall area under the receiver operating characteristic curve (AUROC). Ties were broken by the site-summary score $0.85\bar{A} + 0.15A_{\min} - 0.02s_A$, where $\bar{A}$, $A_{\min}$, and s_A denote the mean, minimum, and standard deviation of AUROC across valid sites. The final model used probabilities for records excluded from each base model's fit, rather than fitted training values.

The database `CreationTime` field was converted to a normalized ordinal, centered within training sites by the median and median absolute deviation, and expanded with two piecewise-linear terms at its tertiles. Global training statistics were used for an unseen site. These label-free transformations were estimated once on the capped training matrix before the final leave-one-site-out evaluation, so a held-out training site's time distribution informed the transformation. The final model contained neither site indicators nor a variable designed to remove site-specific offsets. Two auxiliary feature groups were computed but not used. Within each fold, site-specific models fitted one unweighted boosting model per training site having at least 20 samples and both classes, then averaged their predictions for the excluded site. After all cross-validated predictions had been produced, a single non-nested search tested 0–30% weights for these site-specific models in 5% steps. Candidate weights were compared, in order, by overall AUROC, the site-summary score, mean-site AUROC, and then pooled-model weight. The blend was used only if mean-site AUROC improved without lowering the site-summary score. It was not a separate stacking input and could contribute only through an earlier saved combination; whether it contributed is unknown.

The program also derived eight summaries from the first 30 minutes of an available electrocardiogram: the mean and standard deviation of retained R–R (RR) intervals, root mean square of successive differences, percentage of successive RR-interval differences exceeding 50 ms, low- and high-frequency spectral energy, their ratio, and sample entropy. Peaks in a 5–15-Hz energy envelope smoothed over

0.15 s exceeded its 85th percentile and were separated by at least 0.3 s; RR intervals outside 300–2,000 ms were rejected. The RR series was interpolated at 4 Hz and then centered by the mean of the interpolated values. Hann-windowed squared Fourier amplitudes were summed over 0.04–0.15 and 0.15–0.40 Hz and stored after $\log(1 + x)$ transformation, while their ratio used the untransformed sums. These are implementation-specific approximations to conventional heart rate variability (HRV) domains [9], not calibrated powers or ectopic-cleaned normal-to-normal measures. Sample entropy used at most the first 300 intervals, $m = 2$, and tolerance 0.2 times their standard deviation [10]. A mismatch between training and inference variable names prevented these measures from varying across patients during inference: the HRV columns varied during stacking training, whereas inference filled the unmatched columns with 0.5. We therefore do not attribute the official discrimination result to HRV. If the final stacking model was evaluated, the constant fill could still shift every log-odds score by a constant and thereby change the binary decisions.

When sufficient time remained, the stacking vector x_i concatenated the available base-model probabilities, variables derived from `CreationTime`, and HRV columns for record i. An L2-penalized, class-balanced logistic regression ($C = 6$, maximum 5,000 iterations) estimated $\Pr(Y_i = 1) = \text{logit}^{-1}(\beta_0 + \sum_j \beta_j x_{ij})$. For at least three base-model probability columns, each record's stacking weight was proportional to $(s_i^2 + 0.01)^{-1}$, where s_i was their within-record standard deviation; weights were normalized to mean one. Three bootstrap refits were averaged within each stacking cross-validation fold and for the final fit. A leave-one-site-out analysis of the fixed base-model prediction matrix selected a threshold from 0.15 to 0.85 by balanced accuracy. This evaluation was not nested around base-model fitting: data from its held-out site could contribute to models that produced its training rows. Thus the procedure was used for model selection, not as a strictly nested assessment of a new site, and it did not optimize the official reward. These steps applied only when the full stacking model completed before the time limit.

2.4. Execution and evaluation

When the organizers ran the submitted entry, the program trained models from the released training data in a newly initialized environment; it did not reuse a prefitted model. Missing channels or failed feature calculations propagated as missing values and were handled by the model-specific procedures described above. Training read outcome fields only from the released training partitions. During validation, the program received organizer-provided PSG, metadata, and algorithmic annotations, but no validation outcomes. Inference used no outcome fields or external patient-level data. File and session identifiers served only to locate the corresponding recordings, and patient identifiers were not predictors. The submitted execution path did not depend on external pretrained weights. The reported large-set metrics therefore describe execution of the submitted training and inference program without access to validation outcomes. Independent performance assessment requires a single official test-set evaluation of the selected, unchanged procedure.

We report every aggregate metric returned for the official large-set validation run. The organizers did not return fitted model files, record-level predictions, confidence intervals, calibration estimates, or subgroup results. The task description states that records were de-identified under the Health Insurance Portability and Accountability Act Safe Harbor standard and shared under Beth Israel Deaconess Medical Center Institutional Review Board protocol 2022P000417, with a waiver of informed consent [2].

3. Results

Table 1 gives the complete organizer-returned feedback for official entry 2029. The age-conditioned AUROC was 0.748. The age-weighted and unconditioned AUROCs were 0.813 and 0.790, respectively, and the area under the precision–recall curve was 0.366. The returned binary metrics were reward 0.022, accuracy 0.382, and F-measure 0.167. No final test result or rank was available at manuscript preparation.

Table 1. Organizer-returned large-set validation feedback for official entry 2029.

Metric	Value
Age-conditioned AUROC	0.748
Age-weighted AUROC	0.813
AUROC	0.790
Area under precision–recall curve	0.366
Prevalence-based reward	0.022
Accuracy	0.382
F-measure	0.167

4. Discussion

The ranking and threshold-dependent metrics answer different questions, and the returned aggregates cannot show whether predicted probabilities match observed risk (calibration). We treat them as preliminary organizer feedback, not as an independent performance estimate. If the final stacking model was evaluated, two implementation details require further testing. Its binary threshold optimized balanced accuracy rather than the official reward. In addition, the HRV feature mismatch replaced varying

training features with constants at inference and could change the yes/no decisions. Threshold adjustment alone preserves ranking; refitting after repairing the HRV path may also change ranking.

The evaluated model remains unknown because the time limit could leave an earlier saved model in place; the ranking metrics exclude only a constant prediction. Variable screening, `CreationTime` transformations, and stacking selection were not repeated within every held-out-site fold. The 1,200-record cap also did not preserve site or outcome proportions. Because database creation time may encode cohort chronology or acquisition practice, it should be removed in a separate test before use at a new site.

The primary metric conditions pairwise comparisons on similar age, but this does not establish equitable performance. Demographic variables entered the base models, while the released feedback contained no ranking performance within subgroups, calibration, or error estimates. The final analysis should therefore report prespecified demographic and acquisition subgroups, missing-channel sensitivity, and tests that remove each feature group and `CreationTime` on data not used for model fitting or selection.

5. Conclusion

Entry 2029 combined signal- and annotation-derived sleep measures. Each site was predicted by base models trained on the other sites. Organizer validation returned an age-conditioned AUROC of 0.748. Because the fitted model files and log were not returned, the score cannot be assigned to the final ensemble or an individual feature group. Final assessment requires a single official test-set evaluation of the selected, unchanged procedure. A future version should use the same variables in training and inference, repeat model selection within each held-out-site fold, and retain the fitted model files.

Data and code availability

The data are described on the official task website [2] and in the Human Sleep Project record [4]. Submitted code is identified by Git commit `3040436e`. The organizers state that functional entries will be released after the official phase in accordance with event rules.

Acknowledgments

The author thanks the Challenge organizers. No conflicts of interest or specific funding are reported.

References

[1] Djonlagic I, Aeschbach D, Harrison SL, Dean D, Yaffe K, Ancoli-Israel S, Stone K, Redline S. Associations between quantitative sleep EEG and subsequent cognitive decline in older women. J Sleep Res 2019;28(3):e12666.

[2] George B. Moody PhysioNet Challenge Organizers. Screening for Cognitive Impairment During Sleep Studies: The George B. Moody PhysioNet Challenge 2026. Available at: `https://moody-challenge.physionet.org/2026/`, 2026. Accessed August 20, 2026.

[3] Wolpert DH. Stacked generalization. Neural Netw 1992; 5(2):241–259.

[4] Sun H, Ganglberger W, Nasiri S, Gupta A, Ghanta M, Moura Junior V, Cash S, Stone K, Zhang Z, Ganjoo G, Nassi TE, Wei R, Meulenbrugge EJ, Au R, Clifford G, Trotti LM, Hwang D, Mignot E, Katwa U, Westover MB. The Human Sleep Project. Brain Data Science Platform, 2023. Version 2.0.

[5] Nasiri S, Ganglberger W, Nassi T, Meulenbrugge EJ, Moura Junior V, Ghanta M, Gupta A, Stone KL, Kjaer MR, Sum-Ping O, Mignot E, Hwang D, Trotti LM, Clifford GD, Katwa U, Sun H, Thomas RJ, Westover MB. CAISR: achieving human-level performance in automated sleep analysis across all clinical sleep metrics. Sleep 2025; 48(8):zsaf134.

[6] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189–1232.

[7] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42.

[8] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 2005; 67(2):301–320.

[9] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation and clinical use. Circulation 1996; 93(5):1043–1065.

[10] Richman JS, Moorman JR. Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol 2000;278(6):H2039–H2049.

Address for correspondence:

Yuxuan Xu
College of Computing
Georgia Institute of Technology
801 Atlantic Drive
Atlanta, GA 30332-0280, USA
Email: yxu923@gatech.edu