Multimodal Late-Fusion Prediction from Polysomnography Using Metadata and Timing-Aware Signal Features

Xin Li[1,2,3], Ekenedirichukwu N. Obianom[6,7], Abdulhamed M. Jasim[1,5], Shamsu Idris Abdullahi[1], Fan Feng[1], Abdulmalik, Koya[2,4],G André Ng[2,3,4] and , Noor N. Qaqos[1,8]

[1]School of Engineering, University of Leicester, Leicester, UK; [2]Leicester National Institute for Health and Care Biomedical Research Centre, Leicester, UK; [3]Leicester British Heart Foundation Centre of Research Excellence, Leicester, UK; [4]Division of Cardiovascular Sciences, University of Leicester, Leicester, UK; [5]College of Electronics Engineering, Ninevah University, Mosul, Iraq; [6]iiTech, Infection Innovation Consortium, Liverpool, UK; [7]Vector Biology, Liverpool School of Tropical Medicine, Liverpool, UK; [8] Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq

Abstract

For the PhysioNet Challenge 2026, Leicester Fox team developed a mul-timodal late-fusion pipeline for patient-level prediction from polysomnogra-phy (PSG) and structured clinical metadata, designed to predict future diag-noses of cognitive impairment. We developed a site-aware model combining gradient-boosting and XGBoost metadata branches with an ExtraTrees signal branch derived from multiscale physiological windows and automated sleep annotations. Locally, age-conditioned area under the receiver operating characteristic curve (AUROC) was 0.7027 in five-fold out-of-fold and 0.6728 in leave-one-site-out validation. On official hidden validation, submission 2718 achieved age-conditioned AUROC 0.775, and public rank 15 out of 514 entries.

1. Introduction

Changes in sleep architecture can precede clinically recognized cognitive impairment. Prospective studies have associated reduced rapid eye movement (REM) sleep with incident dementia [1], sleep-disordered breathing and hypoxia with later mild cognitive impairment or dementia [2], and short sleep duration in midlife with dementia incidence over long follow-up [3]. These observations motivate analysis of the complete overnight record rather than a single summary index. However, polysomnography (PSG) is heterogeneous across hospitals: channel availability, recording duration, annotation practice, patient age, prevalence, and clinical follow-up differ by site. A model can therefore rank well within one cohort yet fail after transfer to a previously unseen source.

The Complete Artificial Intelligence Sleep Report (CAISR) provides standardized sleep-stage, arousal, respiratory-event, and limb-movement annotations from heterogeneous PSG signals [4]. For the 2026 George B. Moody PhysioNet Challenge, algorithms predict a future cognitive impairment diagnosis from PSG and metadata. The primary metric is age-conditioned area under the receiver operating characteristic curve (AUROC), defined by positive-negative pairs whose ages differ by no more than two years; Reward separately evaluates binary decisions using age-dependent prevalence [5].

Team Leicester Fox therefore developed one integrated pipeline combining metadata, multiscale physiological descriptors, full-night CAISR summaries, and bounded pairwise ranking. Continuous probabilities incorporate the age-matched corrections, whereas binary predictions use the base ensemble probability and an adaptive threshold. Figure 1 connects representative raw PSG signals to the exact engineered feature families used by the final submitted model.

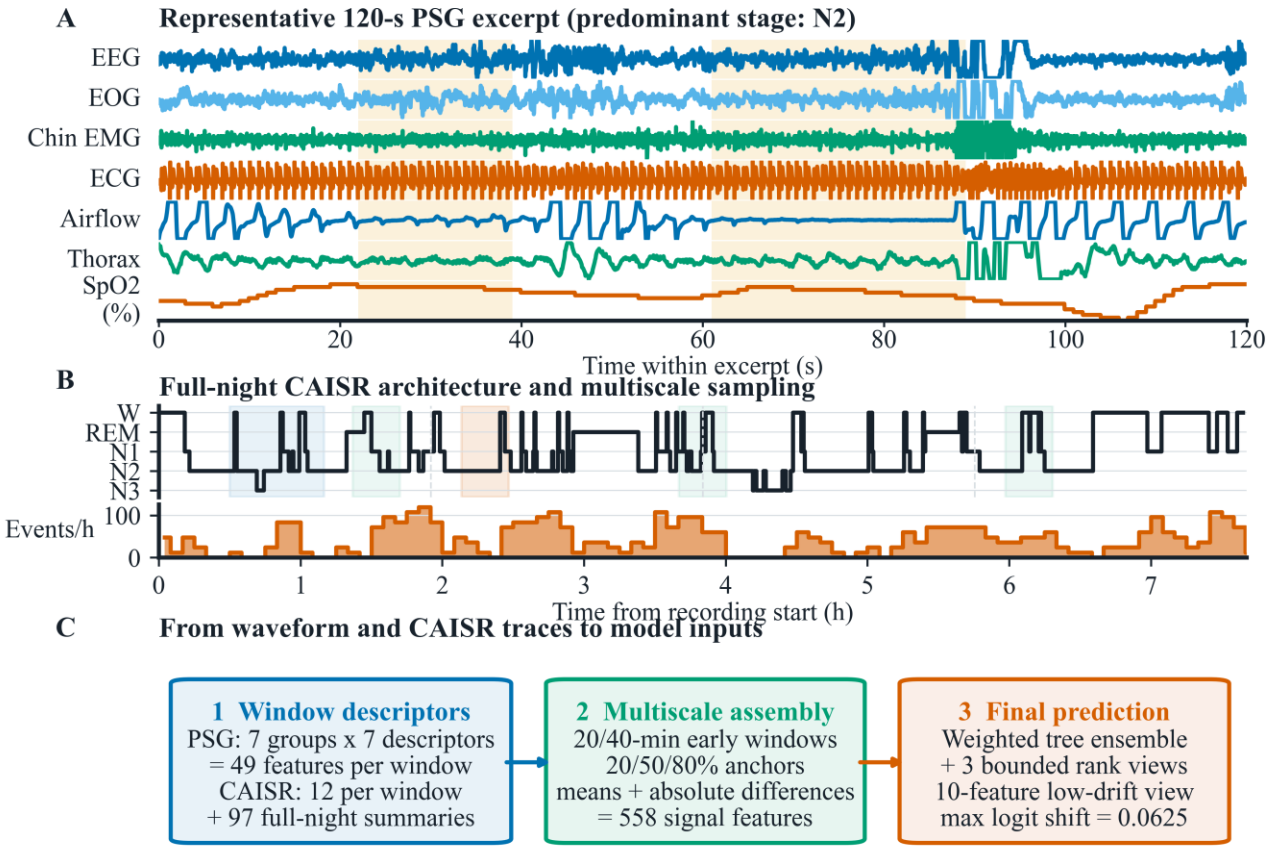

Figure 1. PSG-to-feature pipeline. (A) Anonymized 120-s excerpt; gold shading marks CAISR respiratory events. (B) Full-night hypnogram, event rate, and sampling windows; blue, green, and orange shading denotes early windows, 20/50/80% anchors, and the 30% anchor. (C) Conversion to 558 multiscale features and bounded ranking inputs. Traces are robust-scaled for display only.

2. Methods

2.1. Data, outcome, and preprocessing

The official large set contained 6,600 PSG records: 5,139 from S0001, 319 from I0002, and 1,142 from I0006 [5]. The outcome required at least two qualifying cognitive-impairment diagnoses, with the first one to six years after PSG; 498 records were positive (7.55%). CAISR features were available for 6,530 records. Temporal feature selection used 5,440 labeled records after excluding the 1,103-record local proxy.

Channel names were normalized with a rule table and, when possible, bipolar EEG, EOG, chin EMG, and leg derivations were reconstructed. For the first available channel in each of seven groups (EEG, EOG, chin, leg, ECG, respiratory, and oxygen saturation), we calculated standard deviation, mean absolute value, zero-crossing rate, root mean square, variance, and Hjorth mobility and complexity [6]. Missing channel groups were represented by zero-filled blocks and availability indicators rather than causing record exclusion. Figure 1A shows an anonymized example; robust display scaling was not used during feature computation.

Physiological and CAISR features were computed in 20- and 40-minute windows beginning 30 minutes after recording start, and in 20-minute windows centered at 20%, 50%, and 80% of each usable night. Their means and absolute differences were combined with 97 full-night CAISR descriptors. These included stage fractions, sleep efficiency and fragmentation, stage-transition rate, sleep-onset and REM latency, wake after sleep onset, stage-bout statistics, quartile-specific stage fractions, respiratory/arousal/limb-event rates, and CAISR confidence summaries (Figure 1B,C). The final signal vector contained 558 values. Three augmented training copies used modest feature noise, row-scale jitter, sparse feature dropout, and dropout of complete window, anchor, or full-night blocks.

2.2. Weighted tree ensemble

The base probability was a fixed mean of gradient boosting [7] (300 trees, depth 3), XGBoost [8] (500 trees, depth 3), and ExtraTrees [9] (800 trees, minimum leaf size 4). Metadata branches used visit, calendar, session, site, and missingness fields; training-only follow-up fields were masked stochastically, and Time to Event was excluded. Ensemble weights were 0.400, 0.1833, and 0.4167, respectively.

2.3. Bounded age-matched ranking

The ranking module optimized a pairwise logistic objective related to established learning-to-rank methods [10]. Within each site, positive and negative records were paired only when their ages differed by at most two years. Median-imputed features and missingness indicators were standardized, projected by principal component analysis (PCA), and fitted with an L2-regularized linear score. Four deterministic PCA and pair-subsampling seeds were averaged within each feature view.

The final model combined three complementary rank views. A selected 147-feature physiological view used 32 components, weight 0.075, and clipping at +/-4. Two 32-component views of 220 temporal CAISR features were blended 0.60/0.40, clipped at +/-3, and weighted 0.10. A ten-feature low-drift view retained REM, N1, N2, and N3 fractions and respiratory-event indices from the full night, quartiles, and a 30% anchor. Directions agreed in 89% of 20 half-cohort and site checks.

The low-drift PCA10 scores used four seeds and were clipped individually to +/-3 and in aggregate to +/-0.5. The final probability was sigmoid(logit(p0) + 0.075 cs + 0.10 ct + 0.125 cl), where p0 is the base probability and cs, ct, and cl are the signal, temporal, and low-drift scores. Thus the low-drift logit shift was at most 0.0625. Binary prediction instead compared p0 with an adaptive site-by-age threshold targeting twice the training prevalence, bounded at 5-30%.

2.4. Validation and drift controls

We evaluated age-conditioned AUROC on a 1,103-record local proxy with five-fold out-of-fold (OOF) validation and leave-one-site-out (LOSO) validation. Temporal feature discovery did not use these proxy labels. Features and correction settings were retained only when positive in all eight seed/scheme checks and stable by site. In an unlabeled I0004/I0007 audit, 19 of 20 records were usable; maximum absolute rank shift was 0.2855 logit and the site-mean gap was 0.0979.

The organizers trained the selected code on the full large set and evaluated it on hidden validation. Public leaderboard rank refers to entries, not final team rank, and hidden test labels were unavailable. Models were implemented in Python with scikit-learn and XGBoost.

3. Results

The final model achieved five-fold OOF age-conditioned AUROC 0.702673 and LOSO age-conditioned AUROC 0.672762 on the 1,103-record local proxy (Table 1 and Figure 2A). The low-drift component contributed positively in all eight seed/scheme checks, and selected feature directions agreed in 89% of 20 half-cohort and site checks.

On official hidden validation, Leicester Fox submission 2718 achieved age-conditioned AUROC 0.775 and Reward 0.222 (Table 2 and Figure 2B). The organizer email also reported age-weighted AUROC 0.826, AUROC 0.783, AUPRC 0.459, accuracy 0.759, and F-measure 0.250. This was the team's best official age-conditioned AUROC and occupied public entry rank 15 on 25 August 2026. Hidden validation was a distinct undisclosed source. Because rank correction affected continuous output only, a strong AUROC did not imply a similarly high Reward.

Table 1. Final-model performance. Official hidden validation is a distinct cohort.

Evaluation	n	Age-cond. AUROC	Reward
5-fold OOF	1,103	0.703	-
Leave-one-site-out	1,103	0.673	-
Official hidden	undisclosed	0.775	0.222

Table 2. Organizer-reported hidden-validation metrics for submission 2718.

Metric	Value	Output
Age-conditioned AUROC	0.775	Continuous
Age-weighted AUROC	0.826	Continuous
AUROC	0.783	Continuous
AUPRC	0.459	Continuous
Reward	0.222	Binary
Accuracy	0.759	Binary
F-measure	0.250	Binary

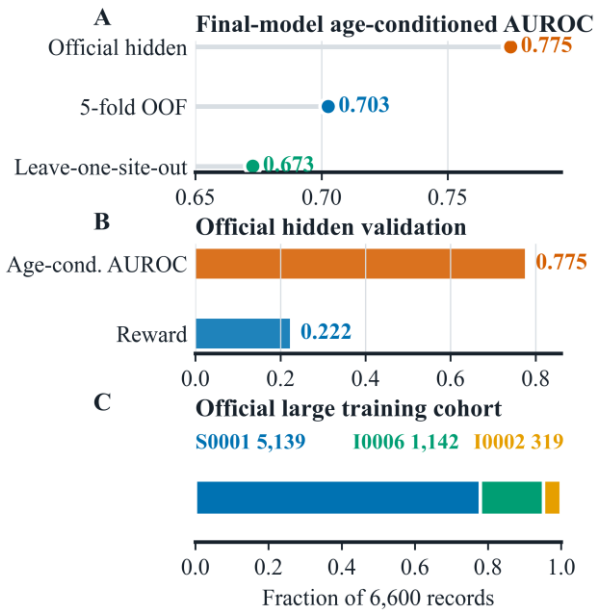

Figure 2. Final-model results. (A) Local OOF, LOSO, and official hidden age-conditioned AUROC. (B) Official AUROC and Reward. (C) Training-set composition. Local and hidden cohorts differ.

4. Discussion

One bounded, site-aware pipeline supported age-conditioned discrimination locally and on hidden validation. Same-site, age-matched pairs reduce the incentive to learn broad age or site differences, while temporal CAISR summaries preserve REM, NREM-depth, and respiratory variation that whole-night averages may dilute. The ten-feature view and clipping constrained the interpretable correction without replacing the weighted tree ensemble.

AUROC and Reward should be interpreted as different tasks. Age-conditioned AUROC depends only on the relative order of continuous probabilities for similarly aged pairs; strictly monotonic calibration leaves that order unchanged. Reward uses binary predictions and age-specific prevalence, making it sensitive to threshold placement under prevalence shift. Here, rankers changed the continuous probability, whereas binary output remained on the base-threshold path. Rank corrections therefore could not directly improve Reward. Future thresholds require strictly nested, site-held-out calibration.

Several limitations constrain interpretation. The 498 positives limit selection within age-site strata, and the 1,103-record proxy remains internal. Masked follow-up fields may encode ascertainment, and CAISR errors can propagate into engineered features [4]. Hidden-test results and confidence intervals were unavailable; the preliminary entry rank does not establish clinical utility.

5. Conclusion

A bounded age-matched model integrating multiscale PSG and temporal CAISR features achieved local age-conditioned AUROC 0.7027/0.6728 and official hidden-validation AUROC 0.775. Reward remained a separate thresholding problem; hidden-test and external validation are required before clinical use.

Acknowledgments

GAN and Xin Li are Supported by the British Heart Foundation Research Excellence Award (RE/24/130031) and Medical Research Council Biomedical Catalyst Developmental Pathway Funding Scheme (MR/S037306/1). Higher Committee for Education Development (HCED), Iraq.

References

[1] Pase MP, Himali JJ, Grima NA, et al. Sleep architecture and the risk of incident dementia in the community. Neurology 2017;89:1244-50.
[2] Yaffe K, Laffan AM, Harrison SL, et al. Sleep-disordered breathing, hypoxia, and risk of mild cognitive impairment and dementia in older women. JAMA 2011;306:613-9.
[3] Sabia S, Fayosse A, Dumurgier J, et al. Association of sleep duration in middle and old age with incidence of dementia. Nat Commun 2021;12:2289.
[4] Nasiri S, Ganglberger W, Nassi T, et al. CAISR: automated sleep analysis across clinical sleep metrics. Sleep 2025;48:zsaf134.
[5] George B. Moody PhysioNet Challenge Organizers. Screening for cognitive impairment during sleep studies: 2026 Challenge. https://moody-challenge.physionet.org/2026/ (accessed 25 August 2026).
[6] Hjorth B. EEG analysis based on time domain properties. Electroencephalogr Clin Neurophysiol 1970;29:306-10.
[7] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189-232.
[8] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min 2016:785-94.
[9] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63:3-42.
[10] Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent. Proc 22nd Int Conf Mach Learn 2005:89-96.

Address for correspondence:
Xin Li
University of Leicester, UK, LE1 7RH
Xin.li@leicester.ac.uk