SIREN: A Site-Invariant Rank Ensemble for Predicting Future Cognitive Impairment from Polysomnography

María Muñoz Pérez[1*], Alejandro Cañada Hinojosa[2*], Lorena Gallego-Viñarás[1], Gonzalo R. Ríos-Muñoz[1], Arrate Muñoz-Barrutia[1]

[1] Universidad Carlos III de Madrid, Madrid, Spain
[2] Imperial College London, London, United Kingdom

[*] These authors contributed equally

Abstract

As part of the George B. Moody PhysioNet Challenge 2026, we (team NeuroUC3M) developed SIREN (Site-Invariant Rank Ensemble), an algorithm to rank patients by their risk of a cognitive-impairment diagnosis within one to six years from a single overnight polysomnogram.

A polysomnogram encodes rich information through many channels, but it also encodes montage conventions and institutional hardware biases, making generalisation to unseen institutions difficult. We therefore enforce site invariance by construction using within-recording percentiles, rank normalisation within institutions, and contrastive augmentations to imitate real-world site discrepancies. SIREN is formed by two branches: an ensemble of logistic, gradient-boosting and LambdaRank heads over 1031 polysomnographic descriptors across five feature views, alongside six frozen self-supervised convolutional encoders over raw electroencephalogram read out by a pairwise attention head. Predictions are then combined via rank averaging. A site probe quantifies the invariance: a classifier separating held-out institutions from the training cohort reaches 0.97 *AUC on the raw descriptors but only* 0.33 *after per-site rank normalisation. On the hidden validation institution, SIREN achieved an age-conditioned AUROC of* 0.755*, placing our team 6th of 94 teams on the official-phase leaderboard. Test set results were not released at the time of writing.*

1. Introduction

Dementia affects over 57 million people worldwide and is projected to nearly triple by 2050 [1]. Diagnosis typically occurs late, once cognitive impairment is clinically evident and substantial neurodegeneration has occurred. Sleep alterations, however, manifest much earlier than symptoms, for example, slow-oscillation–spindle uncoupling [2], slow-wave sleep disruption linked to amyloid-β accumulation [3], or REM sleep without atonia [4]. Polysomnography (PSG) thus offers a promising window for the early detection of neurodegenerative disease [5].

The George B. Moody PhysioNet Challenge 2026 [6] invited teams to develop automated algorithms for predicting future cognitive impairment within one to six years from a single PSG. The primary difficulty is cross-site generalisation: recordings reflect institutional hardware and montage protocols just as strongly as underlying pathology, and the validation and test cohorts each come from an institution absent from the training set.

We present SIREN (Site-Invariant Rank Ensemble). Our contributions are threefold: (i) a two-branch architecture that reads the same night from two perspectives, combined as ranks; (ii) a preprocessing scheme that enforces site invariance at two scales, per night and per site, plus augmentations that imitate measured inter-site discrepancies; and (iii) a site probe that quantifies residual institutional information in a representation, used as a design criterion rather than only as a diagnostic.

2. Methods

SIREN reads the same night twice (Figure 1): *Branch A* over descriptors summarising the whole polysomnogram, *Branch B* over raw electroencephalogram (EEG) encoded without labels, combined by averaging ranks.

2.1. Data and preprocessing

The Challenge corpus [6] is drawn from the Human Sleep Project and distributes 6600 training nights from three US institutions, with hidden validation (I0004) and test (I0007) cohorts from two further ones. Labels come from ICD-9/10 codes as a cognitive-impairment diagnosis one to six years after the PSG; patients diagnosed outside that window, or with under six years of follow-up, are excluded from scoring, so the negative class is survivorship-selected towards sustained clinical contact. We keep 6582

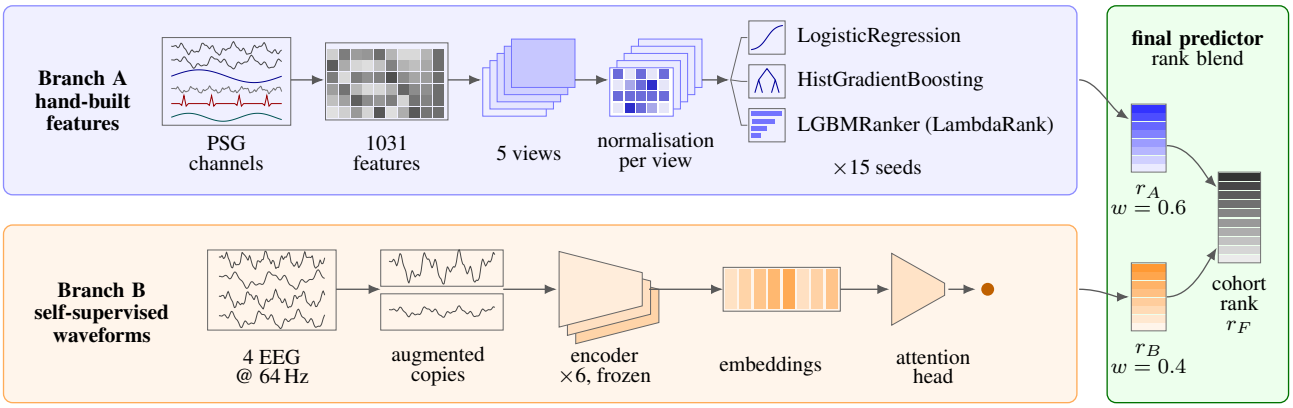

Figure 1. The SIREN pipeline. *Branch A* takes the whole polysomnogram and summarises it as 1031 descriptors. The matrix is split into five feature views, and each view is then normalised and fitted by three heads over 15 seeds, 225 models in all. *Branch B* sees the EEG only, four derivations at 64 Hz, from which six convolutional encoders are pretrained without labels on augmented copies of a 30 s epoch, then frozen and read out by a 258-parameter attention head. The branches are combined by averaging ranks.

nights: S0001 (5123), I0006 (1142) and I0002 (317), discarding 18 recordings with fewer than 20 scorable 30 s epochs, against a median night of 7.5 hours. Prevalence is 7.6%, within the 5–15% reported for the hidden cohorts. Staging and events come from the supplied CAISR annotations, the only staging available for all partitions.

2.1.1. Hand-built features

Branch A reads the whole montage. Spectra are computed over six EEG derivations: central (C3–M2, C4–M1), occipital (O1–M2, O2–M1) and frontal (F3–M2, F4–M1), each resampled to 128 Hz before any work. The electro-oculogram, chin electromyogram, electrocardiogram, oximetry and airflow supply the rest, giving 1031 features per night (Figure 2). There are four groups of descriptors. *Spectral*: 598 descriptors from Welch spectra of 4 s windows capped at 45 Hz, recomputed within each of six stage groupings (W, N2, N3, REM, NREM and all-stage; N1 excluded, its fragmentation gives unstable spectra) covering relative band powers, their ratios (to cancel the amplifier gain), spectral entropy, edge frequencies, aperiodic slope, and peak alpha and sigma frequencies. Each grouping holds 101 and all-stage 93, lacking stage-specific features. *Asymmetry and gradient*: 36 columns, interhemispheric asymmetry for N2 and REM and the anterior–posterior gradient for N2, N3 and REM. *Non-spectral*: 392 carrying no stage tag including spindle density and frequency, slow-oscillation slope and slow-oscillation–spindle coupling phase [2], coherence, permutation entropy, a depth index, heart-rate variability, oxygen desaturation and chin atonia [4]. *Demographic*: five descriptors, being sex, body mass index and three race indicators. Age is not included since the evaluation metric is age-conditioned and keeping it only invites an age detector. These five are the only columns copied into every view, so that none is left without a patient descriptor.

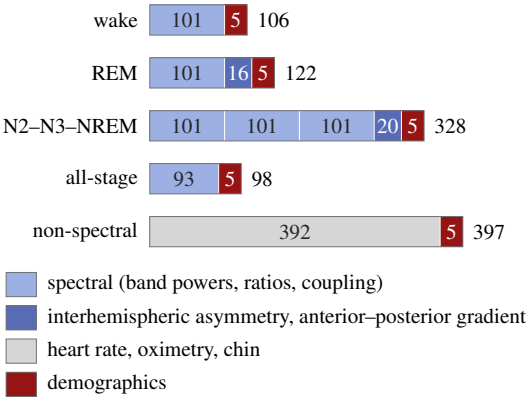

Figure 2. The five views obtained from the 1031-column matrix. Every stage grouping carries a spectral block of 101 columns, or 93 for all-stage, which lacks the stage-specific extras. Interhemispheric asymmetry is computed only for N2 and REM and the anterior-posterior gradient only for N2, N3 and REM. Demographics columns repeat for all views, totalling 1051 features being their union 1031.

Site invariance is then built in by two normalisations at two scales. *Per night*, event detectors use within-recording percentiles. For example, a spindle is the top 5% of sigma power in a given recording rather than an absolute microvolt threshold, so the detector cannot depend on the amplifier. *Per site*, every column of the finished matrix is replaced by its percentile rank among that institution's recordings, on a shared 256-level grid so that cohort size

carries no information.

2.1.2. Raw EEG and augmentations

Branch B sees the EEG alone, and only the four central and occipital derivations, at 64 Hz, cached as 128 stage-stratified 30 s epochs per night. A Nyquist frequency of 32 Hz keeps slow oscillations, K-complexes, spindles and beta intact, and discards the 30–45 Hz band, where what reaches the recording is dominated by each site's anti-alias filtering rather than by the brain. Branch A retains that band, since its ratios and per-site rank normalisation absorb the filter-induced offset. Each recording is scaled at cache time by its robust per-channel standard deviation, so amplifier gain never reaches the network.

To enforce site invariance, we design data augmentations. Each transform simulates an empirical discrepancy we measured across the training sites: per-channel amplitude scaling of 0.5–2$\times$ for amplifier gain, independent 20% channel dropout for unrecorded channels, polarity flips for montage sign conventions, pink noise for the amplifier noise floor, and time shifts for epoch alignment.

2.2. Branch A: ensemble architecture

Branch A has three learners: an L_2-regularised logistic regression, a histogram gradient-boosting classifier, and a LambdaRank ranker grouped into five-year age bands. The ranker directly optimises the ordering of age-matched patients to align with the Challenge evaluation metric.

Each of the five views in Figure 2 is trained independently across all three learners over 15 random seeds (using 70% subsamples), giving $5 \times 3 \times 15 = 225$ models, so that no single feature set or model family dominates. The 225 are combined by averaging percentile ranks rather than raw outputs, which share no common scale and, for LambdaRank, are not probabilities at all.

2.3. Branch B: the raw-waveform encoder

The encoder is pretrained in a self-supervised manner via contrastive learning. Two augmented views of the same 30 s epoch are pulled together while views from different recordings are pushed apart using an NT-Xent loss [7], with sleep-stage prediction as an auxiliary task. No diagnostic labels are involved, so every recording takes part, including unlabelled ones.

The encoder itself is a five-block 1-D convolutional network of 192 k parameters, normalised with GroupNorm. BatchNorm is avoided so that a prediction can never depend on which other recordings happen to share its batch. The encoder is then frozen, and a small head weighs the night's epochs and reduces them to one score, trained on age-matched pairwise ordering. Keeping the head small limits overfitting to site-specific artifacts. Branch B is additionally an ensemble of six encoders differing in augmentation strength and auxiliary-loss weight rather than in seed, again combined by averaging percentile ranks.

2.4. Combining the two branches

To align with the evaluation metric, the branches are combined as rankings. A record's percentile ranks r_A and r_B are averaged with weight $w = 0.4$ on Branch B, giving $r_F = 0.6\, r_A + 0.4\, r_B$, and the cohort is re-sorted by r_F. Keeping $w < 0.5$ leaves Branch A in charge of the global ordering, so Branch B settles near ties while widely separated pairs remain untouched. Records with unusable waveforms retain r_A, so Branch B is an improvement and never a dependency.

2.5. Training and validation protocol

Every numerical result in this paper is *leave-one-site-out* (LOSO): fit on two institutions, score on the third. We steer by the mean of the two folds holding out I0002 and I0006, not the one holding out S0001, because the hidden institutions are small and single whereas S0001 alone is 78% of the corpus. We initially tried random k-fold cross-validation but it mixes institutions across train and test, and proved badly overoptimistic against the hidden site.

Within a fold, Branch A is fitted on the larger of the two training institutions rather than on both pooled, since pooling mixes incompatible measurement conditions and lets the majority institution set the scale for the other. Branch B pretrains on every usable night in the fold, labelled or not, and labels enter only through the final head. The blend weight is chosen the same way: to score a fold, w is selected on the other two, so no fold helps choose the weight it is then judged under.

3. Results

Entries are scored by an age-conditioned AUROC: over every positive–negative pair whose ages differ by at most two years, the fraction the model orders correctly. Every stage of our pipeline therefore works in ranks. The Challenge additionally reports a prevalence-based reward metric that depends on a binary decision; since SIREN is a ranker, our binary output is a thresholded rank and we did not optimise for it.

Table 1 reports LOSO results. Neither branch dominates, the blend leads on either mean. On the small-site mean, demographics without age reach 0.582 and age alone 0.525, against our 0.7713. On the hidden validation institution (I0004) the submitted blend scored 0.755, placing us 6th of 94 on the official-phase leaderboard; test set results are released after the conference.

Table 1. LOSO age-conditioned AUROC over three seeds. The small-site mean averages the I0002 and I0006 folds.

Held-out site (n)	Branch A	Branch B	Blend
I0002 (317)	0.7485	**0.7962**	0.7829
I0006 (1142)	0.7497	0.7310	**0.7598**
S0001 (5123)	**0.6380**	0.6121	0.6404
Small-site mean	0.7491	0.7636	**0.7713**
3-site mean	0.7121	0.7131	**0.7277**

To check that our invariance mechanisms work we train a *site probe*: a classifier separating the two hidden sources from training records, given one representation alone. The Challenge supplies twenty unlabelled examples from I0004 and I0007, too few to score but enough to measure shift, pooled here into a single held-out class (Figure 3).

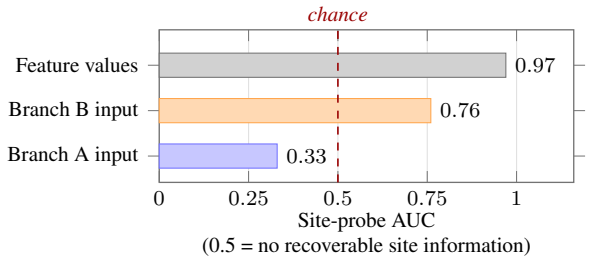

Figure 3. How easily a classifier separates the 20 hidden-institution records from training records, given the 1031 features, the encoder's embeddings, and the features after per-site rank normalisation.

4. Discussion and conclusions

Every component of SIREN was chosen for generalisation, and the site probe indicates it worked: the 1031 descriptors identify the institution at 0.97 AUC, Branch B's learned embeddings at 0.76, and the same descriptors after per-site rank normalisation at 0.33 — no better than chance in either direction. The blend scored 0.755 on the hidden validation institution.

Several limitations qualify this. The main bottleneck is mild cognitive impairment (MCI), 31% of the positives (n = 152), where we reach 0.625 against 0.735–0.759 across the dementia subtypes. The subgroups are small, but the gap holds in sign across all three folds. MCI is an earlier and more heterogeneous endpoint, so one binary label asks a single ranker to order two rather different transitions at once. Ranking by expected time to diagnosis would separate them without subtype labels the hidden sites lack. SIREN is also transductive, ranking a cohort rather than scoring a patient, so a new institution must supply a reference set before anyone in it can be ranked. Twenty records suffice to detect a gross fingerprint but not to certify its absence, and the negative class requires six years of sustained follow-up, a selection no deployed model would encounter.

These results support the premise the Challenge was built on: a single PSG carries a usable signal about cognitive decline years before diagnosis. What limits such a model in practice is not the physiology but the recording, since it describes an institution as plainly as it describes a pathology. Making that description site-invariant is the central problem in carrying this into a clinic.

Code will be released at https://github.com/mmunozpz/PhysionetChallenge2026-NeuroUC3M once the Challenge concludes.

Acknowledgments

This work was supported by projects PI22-01619, PID2025-175513OB-I00, PID2023-152631OB-I00, AIA2025-164165, FLAMA-CM-UC3M, MAGERIT-CM (TEC-2024/COM-44) and SMART-CM (SYG-2024/SAL-GL-619).

References

[1] GBD 2019 Dementia Forecasting Collaborators. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. Lancet Public Health 2022;7(2):e105–e125.

[2] Helfrich RF, Mander BA, Jagust WJ, Knight RT, Walker MP. Old brains come uncoupled in sleep: slow wave-spindle synchrony, brain atrophy, and forgetting. Neuron 2018; 97(1):221–230.

[3] Ju YES, Lucey BP, Holtzman DM. Sleep and Alzheimer disease pathology—a bidirectional relationship. Nat Rev Neurol 2014;10(2):115–119.

[4] Postuma RB, Iranzo A, Hu M, Högl B, Boeve BF, et al. Risk and predictors of dementia and parkinsonism in idiopathic REM sleep behaviour disorder: a multicentre study. Brain 2019;142(3):744–759.

[5] Mander BA, Winer JR, Jagust WJ, Walker MP. Sleep: A novel mechanistic pathway, biomarker, and treatment target in the pathology of Alzheimer's disease? Trends in Neurosciences 2016;39(8):552–566.

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

[7] Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In Proc Int Conf Mach Learn. 2020; 1597–1607.

Address for correspondence:

María Muñoz Pérez

30 Av. de la Universidad, 30, 28911 Leganés, Madrid, Spain

maria.munozp@alumnos.uc3m.es