Cross-Site Generalization of Interpretable Sleep Features for Predicting Future Cognitive Impairment

S. A. I. Shouborno*, Bashima Islam

Worcester Polytechnic Institute
Worcester, MA, USA

Aims. We aimed to identify which polysomnographic descriptors of future cognitive decline remain predictive when training and evaluation sites differ.
Methods. From 622 overnight recordings across three clinical sites (36, 102, and 484 subjects), 274 descriptors were extracted spanning seven families: probabilistic stage and event labels, stage transition and bout statistics, per-stage EEG spectral power, slow-wave and spindle morphology, heart-rate variability, SpO_2 indices, and EEG complexity (using YASA, NeuroKit2, and AntroPy). As a domain-generalization step, descriptors whose site-prediction AUC exceeded 0.65 or whose class-association sign flipped across sites were discarded (137 retained). Three L_2-regularized logistic-regression configurations (C=0.005, median imputation) were evaluated under leave-one-site-out (LOSO) cross-validation on the area under the receiver operating characteristic curve (AUROC): a dual-model ensemble combining site-specific feature subsets, a single worst-site-robust model whose 28 features came from greedy forward selection against worst-site AUROC, and a four-model per-site ensemble.
Results. Our dual-model submission (team `bashlab_wpi`) reached a LOSO mean AUROC of 0.740 but only 0.585 on its worst site. On the official validation set it scored 0.674. The single worst-robust model raised the weakest fold to 0.672 (from 0.585 in the dual model) with a LOSO mean of 0.780, though validation fell to 0.644. A four-model per-site ensemble produced a higher LOSO mean (0.805) but a lower worst-site AUROC (0.630) (Table 1).

Approach	LOSO mean	LOSO worst	Validation
Dual-model LR (31+12 features)	0.740	0.585	**0.674**
Single LR, worst-robust (28 features)	**0.780**	**0.672**	0.644
Four-model per-site ensemble	0.805	0.630	—

Table 1. LOSO cross-validation vs. held-out validation AUROC. The gap reflects the cross-site generalization challenge on the held-out validation site.

Conclusion. LOSO cross-validation overestimated held-out performance for all configurations, suggesting that site-prediction filtering alone does not fully resolve cross-site distribution shift. Ongoing work incorporates domain adaptation using unlabeled recordings from the held-out sites to further narrow this gap.