A Benchmark Audit of Site Confounds, Calibration, and Self-Supervision in Cross-Dataset Parkinson's EEG Detection

Edward Lue Chee Lip, Van Mai, Saanvi Neema, Alexander Jameson, Karolina Torbus, and Jithin Suresh

Abstract. Electroencephalography (EEG) is a low-cost, non-invasive candidate biomarker for Parkinson's disease (PD), and deep-learning models report high balanced accuracy on a benchmark that pools four public PD datasets. We show that, in this benchmark, pooled accuracy is strongly influenced by site-associated label imbalance: a diagnostic site-prior null that uses no EEG signal (predicting each dataset's majority class) reaches 0.90 segment-level balanced accuracy, comparable to published models. Evaluating instead with three-site leave-one-dataset-out (LODO) over the both-class PD datasets, fixed-threshold balanced accuracy is substantially degraded. Our analysis suggests that this degradation includes a major threshold/calibration component rather than a complete loss of discriminative information: supervised scores rank PD versus healthy controls on unseen sites at ROC-AUC 0.76 ± 0.03, and a fully deployable decision threshold selected on the training sites reaches 0.64 ± 0.03 balanced accuracy. Finally, under the architecture, data, and scale tested, self-supervised pretraining provides no observed cross-site gain: frozen linear probes on encoders pretrained on the four datasets or on a large disjoint clinical corpus (TUH) reach cross-site AUC of 0.58 and 0.53 respectively, and fine-tuning matches but does not improve over training from scratch, with no data-efficiency advantage. We will release the evaluation tooling and recommend reporting cross-site PD-EEG with LODO, threshold-independent metrics, and an explicit site-prior null.

Keywords: EEG · Parkinson's disease · cross-site generalization · dataset confound · calibration · self-supervised learning.

1 Introduction

Parkinson's disease is the second most common neurodegenerative disorder, and diagnosis still rests on motor signs that appear only after substantial dopaminergic loss. EEG is inexpensive, non-invasive, and widely available, making an EEG marker attractive for screening. Deep-learning models report balanced accuracies near 80–90% on public PD-EEG benchmarks [5,11,12], but a marker is clinically useful only if it holds on patients and equipment the model never saw, and the standard pooled benchmark does not test this.

The pooled protocol holds out subjects but not *sites*: every dataset appears in both training and test folds. Because EEG recordings are identifiable by site

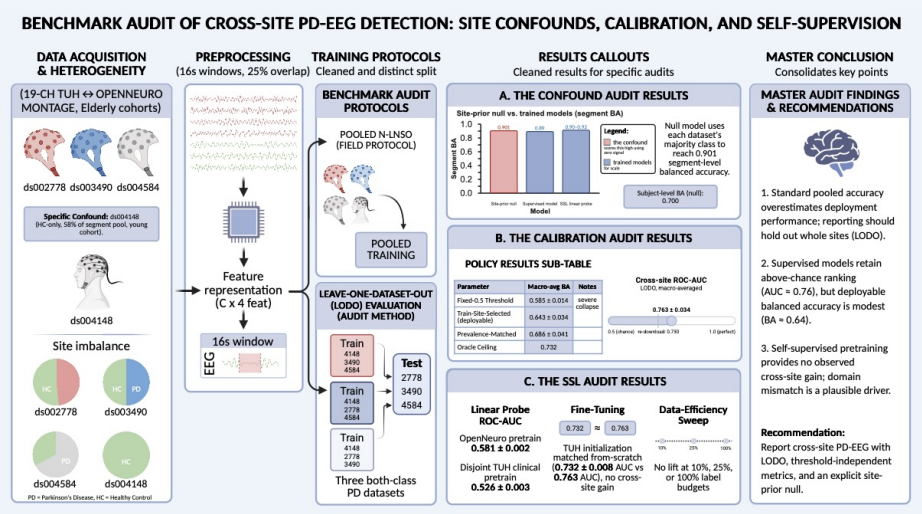

Fig. 1. Benchmark-audit workflow and principal findings. Four public PD–EEG datasets are harmonized to a 19-channel montage, segmented into 16-s windows, and evaluated using pooled N-LNSO and three-site LODO. The audit isolates **(a)** site-label confounding through a no-EEG site-prior null, **(b)** threshold and calibration shift under LODO, and **(c)** transfer from OpenNeuro or TUH VICReg pretraining. The final panel summarizes the resulting cross-site recommendations.

(amplifier, montage, filtering) and each dataset carries a dataset-specific label distribution, high pooled accuracy can reflect dataset recognition rather than disease. Whether such accuracy reflects pathology, and whether it transfers to a genuinely unseen site, has not been directly tested.

We make three contributions. **(1)** We introduce a *site-prior null* and show that the pooled-protocol accuracy is strongly influenced by site effects in this benchmark: a no-EEG baseline is comparable to published models. **(2)** Using three-site LODO over the both-class PD datasets, we separate discrimination from thresholding: supervised scores retain above-chance ranking (AUC ≈ 0.76), while deployable balanced accuracy remains modest and threshold-sensitive. **(3)** We report a benchmark-specific negative result: under our tested architecture, pretraining data, and label budgets, self-supervised pretraining provides no observed cross-site gain across linear-probe, fine-tune, and data-efficiency evaluations. Figure 1 summarizes the pipeline.

2 Related Work

Evaluating PD-EEG across datasets. Deep models for PD-EEG are typically evaluated by pooling several public datasets and cross-validating across subjects [5]; this holds out subjects but not the dataset of origin. Segment-level leakage (windows from one recording appearing in both train and test) is

known to inflate accuracy and collapses by over twenty points once cross-subject splits are enforced [5], and preprocessing choices alone swing reported accuracy substantially [3]. A parallel line works to harden the protocol itself: nested stratified cross-validation with channel selection reaches 80.6% cross-dataset [11], and population-aware frameworks report up to 94.1% on held-out cohorts as training diversity grows [12]. We extend this line in a different direction: rather than seeking a better cross-site number, we ask what the standard pooled number *measures*, and introduce a no-EEG site-prior null and a calibration decomposition that neither of these works reports.

Self-supervised learning for EEG. Large-scale self-supervised pretraining underpins recent EEG foundation models such as LaBraM [7], BIOT [16], and BENDR [8]; the last of these established the now-standard recipe of pretraining on the large clinical TUH corpus and fine-tuning on a small downstream task. We test this recipe for cross-site PD detection, using VICReg [1] via SelfEEG [4], and report a benchmark-specific negative result. A generative alternative, GAN-based augmentation [10], pursues the same generalization goal we test here.

Calibration under domain shift. Modern networks are systematically miscalibrated, and cheap post-hoc fixes such as temperature scaling are well established [6]. We examine whether the apparent cross-site degradation of PD-EEG models reflects threshold/calibration shift in addition to weak or noisy signal under domain shift, connecting this literature to a domain where it has not previously been applied.

3 Methods

3.1 Datasets.

We use four public OpenNeuro PD datasets (290 subjects: 140 PD, 150 healthy controls [HC]) and the Temple University Hospital (TUH) EEG corpus [9] as an unlabeled pretraining source. The labeled datasets differ markedly in composition: ds004148 [15] contributes 60 HC and no PD; ds002778 [13] (UC San Diego) 15 PD / 16 HC; ds003490 [2] (UNM 3-Stim) 25 PD / 25 HC; and ds004584 [14] 100 PD / 49 HC. We obtained the labeled data by re-downloading from the public OpenNeuro mirror and reprocessing the resting-state recordings locally; after windowing, ds004148 alone accounts for 8,640 of 14,992 segments (58%), all HC—a composition central to the confound in Section 4.1.

3.2 Preprocessing and channel harmonization.

Recordings are band-pass filtered (1–45 Hz), resampled to 250 Hz, segmented into 16 s windows with 25% overlap, and z-scored per channel within each segment (no statistics shared across the train/test boundary). The original benchmark uses the 29 channels common to all four OpenNeuro datasets; however, measuring channel presence directly on TUH shows only 15 of these 29 carry signal (ten extended 10–10 positions are absent; four appear under legacy names

T3/T4/T5/T6). We add an old-to-new name remap and adopt the **19-channel TUH∩OpenNeuro montage** (verified present in TUH), so a TUH-pretrained encoder never learns channel filters on dead inputs. All cross-site experiments use the 19-channel montage.

3.3 Architecture.

We use the TransformEEG encoder (depthwise convolutional tokenizer → two-layer transformer → adaptive pooling), producing a $C \times 4$ feature, where C is the encoder embedding dimension and 4 the pooled temporal length. A two-layer head is attached for supervised training; a VICReg projector for self-supervised pretraining.

3.4 Evaluation protocols.

Combined nested leave-N-subjects-out (N-LNSO) is the pooled protocol: all four datasets are merged and 10-fold subject-stratified cross-validation holds out subjects; every dataset appears in every fold. **Leave-one-dataset-out (LODO)** holds out one dataset and trains on the other three. We evaluate LODO over the three both-class PD datasets; ds004148 is HC-only and is not a balanced-accuracy test fold. Alongside the pooled protocol we report a **site-prior null**: a diagnostic classifier that predicts the majority class associated with each segment's dataset, using no EEG signal. The null is a pooled-protocol, label-only reference; under LODO the held-out site's label distribution is unavailable at training time, so the null is shown only for reference. LODO metrics are macro-averaged over the three held-out sites and over seeds where available.

3.5 Metrics and calibration policies.

We report subject-level balanced accuracy and **ROC-AUC**; for LODO, segment probabilities are averaged within subject, metrics are computed within each held-out dataset, and sites are macro-averaged. Because a fixed 0.5 threshold miscalibrates on an unseen site, we evaluate: *fixed-0.5*; *train-selected* (the training-site balanced-accuracy optimum, applied unchanged); *prevalence-matched* (positive rate set to held-out prevalence); and *oracle* (maximum held-out balanced accuracy over thresholds). Variability is mean ± standard deviation across seeds after site macro-averaging. For the re-download replication we also report subject-clustered 95% confidence intervals within the observed held-out sites, resampling 230 held-out subjects; with only three held-out sites, these intervals do not estimate uncertainty over future sites.

3.6 Self-supervised pretraining.

We pretrain with VICReg and evaluate a **linear probe** (encoder frozen), **fine-tuning** (end-to-end from the SSL checkpoint), and a **data-efficiency** sweep

(10/25/100% of training subjects). OpenNeuro SSL uses all four datasets unlabeled, so it is target-site-unlabeled under LODO; TUH SSL uses a disjoint subset (298 recordings, 19,883 segments). Supervised/fine-tuning runs use Adam (β_1=0.75), lr 2.5×10^{-4}, 50 epochs.

4 Results

4.1 Pooled-protocol accuracy is strongly influenced by site.

On combined N-LNSO, the supervised model reaches 0.89 median segment-level balanced accuracy and the SSL probe 0.90–0.92. However, the **site-prior null reaches 0.901** (segment) and 0.700 (subject) using no EEG signal (Fig. 2). This null does not require train–test leakage; it exploits the site-label imbalance permitted by the pooled protocol. Because ds004148 is entirely HC and dominates the segment pool while each PD dataset is PD-majority by segment count, predicting the dataset's majority class alone is comparable to the trained models. The pooled metric by itself is therefore not sufficient to distinguish disease discrimination from dataset recognition.

4.2 Cross-site evaluation separates ranking from thresholding.

Under three-site LODO, fixed-0.5 subject-level balanced accuracy is 0.585 ± 0.014—substantially degraded, predicting PD for almost all subjects on an unseen site. Yet the threshold-independent **ROC-AUC is** 0.763 ± 0.034 (Fig. 3): the model ranks PD above HC above chance on sites it never saw. This pattern suggests a substantial thresholding/calibration component, while still leaving open the possibility that the underlying disease signal is weak or noisy under severe site shift. A threshold selected on the training sites and applied unchanged (training-site-selected, fully deployable) improves balanced accuracy to 0.643 ± 0.034; prevalence-matching reaches 0.686 ± 0.041; the held-out oracle is 0.732 (Fig. 2). On an independent re-download of the data, the supervised model reproduces this pattern: a subject-clustered bootstrap within the observed held-out sites (10,000 resamples) gives ROC-AUC 0.730 (95% CI $[0.644, 0.809]$), training-site-selected balanced accuracy 0.660 ($[0.599, 0.721]$), and fixed-0.5 0.586 ($[0.527, 0.642]$). With only three held-out sites the fixed-0.5 and deployable intervals overlap, so we report the gain as a consistent point-estimate improvement—present on two of the three held-out sites—rather than a strictly separated interval. Thus, within these datasets, there is above-chance cross-site discrimination, but performance remains modest and threshold-sensitive.

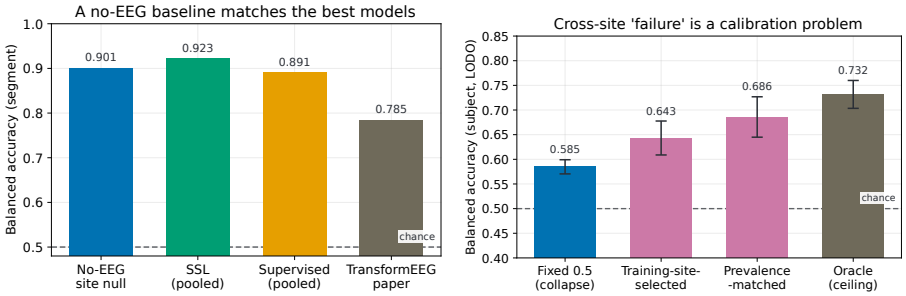

Fig. 2. Left: a no-EEG site-prior null is comparable to the best models on the pooled protocol. **Right:** under LODO, a training-site-selected threshold improves balanced accuracy over the fixed-0.5 operating point. Error bars show standard deviation across seeded runs.

4.3 Self-supervised pretraining provides no observed cross-site gain.

Table 1 focuses the LODO comparison on the SSL question, reporting threshold-independent ranking (ROC-AUC) and deployable balanced accuracy.

Table 1. Cross-site LODO model comparison. Rows are macro-averaged over the three held-out both-class PD datasets.

Training	Encoder source	ROC-AUC	Deployable BA
Supervised	scratch	0.763 ± 0.034	0.643 ± 0.034
Supervised + augment	scratch	0.764 ± 0.029	0.653 ± 0.032
Fine-tune	SSL-OpenNeuro	$0.733^{\dagger}$	$0.610^{\dagger}$
Fine-tune	SSL-TUH	0.732 ± 0.008	0.636 ± 0.025
Linear probe	SSL-OpenNeuro	0.581 ± 0.002	0.501 ± 0.004
Linear probe	SSL-TUH	0.526 ± 0.003	0.527 ± 0.007

Deployable BA uses a training-site-selected threshold. SSL-OpenNeuro includes unlabeled target-site data; SSL-TUH is disjoint. Mean $\pm$ SD across seeds. †single seed.

The fixed-0.5 degradation and no-EEG site-prior null are isolated in Fig. 2 and Fig. 2. We observe no SSL gain across the tested LODO conditions (Fig. 3). *Linear probe:* cross-site AUC is 0.581 ± 0.002 for target-site-unlabeled OpenNeuro pretraining and 0.526 ± 0.003 for disjoint TUH pretraining, versus 0.763 supervised; the larger disjoint corpus gives the lowest AUC, consistent with clinical/resting-state domain mismatch. *Fine-tuning:* TUH initialization matches from-scratch at full labels (0.73 vs 0.76). *Data-efficiency:* across 10/25/100% label budgets, SSL-TUH tracks from-scratch (Fig. 3). *Augmentation:* we observe little cross-site change (0.763 vs 0.764). The SSL-probe negatives are stable across seeds (std $\approx$ 0.002–0.003).

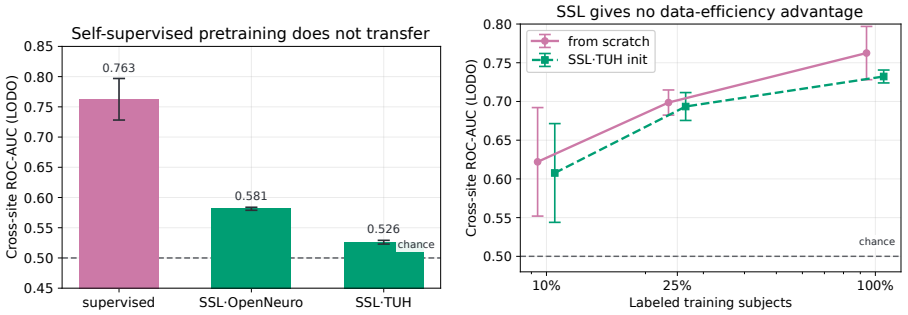

Fig. 3. Left: cross-site ROC-AUC by method: supervised training gives the highest AUC in this benchmark. **Right:** SSL fine-tuning shows no observed data-efficiency advantage in the tested label-budget sweep. Error bars show standard deviation across seeded runs.

5 Discussion

Our results reframe this cross-dataset PD-EEG benchmark. First, the widely reported pooled accuracy is strongly influenced by site effects; the site-prior null is therefore an important baseline, and evaluation should hold out whole sites. Second, the apparent cross-site degradation of supervised models contains a substantial *threshold/calibration* component: a threshold selected on the training sites narrows the gap to the oracle, but the remaining modest accuracy is also compatible with weak or noisy disease signal under site shift. This motivates site-aware calibration (using target-site metadata or a few labeled calibration subjects) alongside better representations. Third, self-supervised pretraining did not deliver a cross-site advantage under our architecture, pretraining corpora, label budgets, and compute scale; one plausible explanation is domain mismatch (clinical vs resting-state EEG), suggesting that *what* one pretrains on may matter as much as scale.

6 Conclusion

Cross-dataset PD-EEG accuracy, as conventionally reported in this benchmark, is strongly influenced by site effects. With leave-one-dataset-out evaluation, threshold-independent metrics, and an explicit site-prior null, supervised scores show above-chance cross-site ranking but modest, threshold-sensitive balanced accuracy; self-supervised pretraining provides no observed improvement under the tested conditions.

Prospects of Application. For screening, PD–EEG models should generalize to unseen clinics. Our benchmark audit suggests a path, not clinical readiness: supervised scores show above-chance cross-site ranking (AUC $\approx$ 0.76) and a training-site threshold reaches balanced accuracy $\approx$ 0.64, while pooled accuracies can overstate deployment performance.

Limitations. This is a benchmark audit, not a clinical deployment study. Four datasets and three held-out sites; a single architecture; small subject counts per held-out site (31/50/149), so the reported point estimates and intervals describe the tested cohort, not generalization to future, unseen sites; SSL pretraining bounded by single-GPU scale (the disjoint TUH subset is ∼20k segments). Cross-site and data-efficiency results use three seeds.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bardes, A., Ponce, J., LeCun, Y.: VICReg: Variance-invariance-covariance regularization for self-supervised learning. In: International Conference on Learning Representations (ICLR) (2022)
2. Cavanagh, J.F., et al.: EEG: 3-stim auditory oddball and rest in Parkinson's (ds003490). OpenNeuro (2021)
3. Del Pup, F., et al.: The more, the better? Evaluating the role of EEG preprocessing for deep learning applications. arXiv preprint arXiv:2311.18046 (2024)
4. Del Pup, F., et al.: SelfEEG: A Python library for self-supervised learning in electroencephalography. Journal of Open Source Software **9**(95), 6224 (2024)
5. Del Pup, F., et al.: TransformEEG: Towards improving model generalizability in deep learning-based EEG Parkinson's disease detection. arXiv preprint arXiv:2507.07622 (2025)
6. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning (ICML) (2017)
7. Jiang, W.B., Zhao, L.M., Lu, B.L.: LaBraM: Large brain model for learning generic representations with tremendous EEG data in BCI. In: International Conference on Learning Representations (ICLR) (2024)
8. Kostas, D., Aroca-Ouellette, S., Rudzicz, F.: BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. Frontiers in Human Neuroscience **15** (2021)
9. Obeid, I., Picone, J.: The Temple University Hospital EEG data corpus. Frontiers in Neuroscience **10**, 196 (2016)
10. others: GEPD: GAN-enhanced generalizable model for EEG-based detection of Parkinson's disease. arXiv preprint arXiv:2508.14074 (2025)
11. Rasmussen, N.R., Rizk, R., Wang, L., Singh, A., Santosh, K.C.: Channel selected stratified nested cross validation for clinically relevant EEG based Parkinson's disease detection. arXiv preprint arXiv:2601.05276 (2025)
12. Rasmussen, N.R., Wang, L., Rizk, R., Akter Pallab, M.R., Stuwart, S., Mancini, M., Singh, A., Santosh, K.C.: Robust and clinically reliable EEG biomarkers: A cross population framework for generalizable Parkinson's disease detection. arXiv preprint arXiv:2604.23933 (2026)
13. Rockhill, A.P., Jackson, N., George, J., Aron, A., Swann, N.C.: UC San Diego resting state EEG data from patients with Parkinson's disease (ds002778). OpenNeuro (2021)
14. Singh, A., et al.: Resting-state EEG in Parkinson's disease (ds004584). OpenNeuro (2023)

15. Wang, Y., et al.: A test-retest resting and cognitive state EEG dataset (ds004148). OpenNeuro (2022)
16. Yang, C., Westover, M.B., Sun, J.: BIOT: Biosignal transformer for cross-data learning in the wild. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)