SPICE: A SUBJECT-PRESERVING COUNTERFACTUAL AUGMENTATION FRAMEWORK FOR EEG DISEASE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG-based disease detection often generalizes poorly to unseen subjects. A key challenge is that each subject contributes recordings with a single diagnostic label, enabling models to exploit subject identity as a shortcut for disease prediction. We propose SPICE (Subject-Preserving Inter-Class Expansion), a counterfactual augmentation framework that challenges this shortcut by generating opposite-class EEG samples while preserving subject-specific characteristics. By providing training subjects with examples of both diagnostic classes, SPICE aims to encourage learning of disease-related features that transfer across individuals. We formulate counterfactual generation as unpaired translation between healthy and disease domains and instantiate the framework using CycleGAN with subject-preservation objectives. These objectives also guide a task-specific selection policy that evaluates subject preservation and diagnostic class change while near-copies are screened out of the training view. Experiments across four EEG datasets demonstrate the potential of subject-preserving inter-class expansion to improve cross-subject disease detection, with benefits that depend on the dataset and classifier.

1 INTRODUCTION

Electroencephalography (EEG) provides a non-invasive measure of neural activity that can reveal disease-related abnormalities, motivating growing interest in deep learning for automated disease detection (Roy et al., 2019; Lucas et al., 2024). However, clinical utility hinges on cross-subject generalization: models must diagnose patients unseen during training. In practice, models that achieve near-perfect accuracy under random train-test splits suffer catastrophic performance drops under strictly subject-disjoint evaluation (Wang et al., 2024b; Brookshire et al., 2024).

This failure stems from a fundamental structural property of diagnostic datasets: the *single-class-per-subject* (SCPS) paradigm (Li et al., 2026). Because clinical conditions are largely static, each participant provides multiple EEG segments but only one diagnostic label (Y). Consequently, the diagnostic label is a deterministic function of subject identity (S). Because EEG signals carry strong, individualized biometric signatures, standard empirical risk minimization allows networks to minimize training loss by learning an identity-to-diagnosis lookup table (Wang et al., 2024b) rather than invariant disease biomarkers. In contrast, in *multi-class-per-subject* (MCPS) tasks (e.g., emotion recognition, motor imagery), each participant exhibits multiple states, ensuring Y cannot be trivially predicted by identifying S (Brookshire et al., 2024; Li et al., 2026).

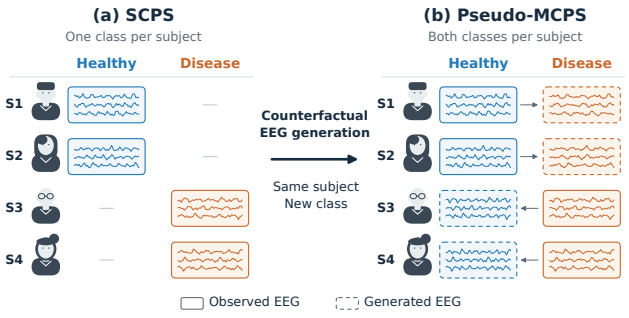


Figure 1: SPICE augments single-class training subjects with opposite-class synthetic counterparts, aiming to retain subject-specific characteristics while introducing within-subject class variation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To challenge this shortcut in SCPS tasks, we introduce **SPICE (Subject-Preserving Inter-Class Expansion)**. SPICE transforms the training distribution from SCPS into a pseudo-MCPS regime by synthesizing counterfactual EEG recordings: generating how a patient's EEG would manifest in the opposite diagnostic state while retaining their underlying subject identity (Figure 1). Because counterfactual ground truth is fundamentally unobservable, we formulate the transformation as unpaired signal translation paired with explicit subject-preservation and disease-conversion objectives. These constraints also govern a targeted sample-selection policy that filters out degenerate translations and trivial copies. By decoupling subject identity from the diagnostic label at the data level, SPICE is designed to encourage models to learn disease features that transfer across individuals.

Our main contributions are summarized as follows:

- **Conceptual Paradigm:** We propose inter-class expansion via counterfactual sample generation as a principled strategy to break identity-based shortcuts in SCPS datasets, expanding the training distribution toward a pseudo-MCPS regime where, for subjects with accepted counterparts, subject identity no longer dictates the diagnostic label.
- **Generative Formulation:** We instantiate this framework as an unpaired signal translation problem using a CycleGAN architecture enhanced with explicit subject-preservation constraints, enabling the synthesis of opposite-class EEG without requiring paired longitudinal patient data.
- **Selection Policy:** We design a quality-filtering mechanism that audits candidate counterfactuals, aiming to select candidates that score as diagnostically converted while remaining subject-similar.
- **Empirical Evaluation:** We evaluate SPICE across four EEG cohorts and multiple models under subject-disjoint evaluation. Comparisons with real-only training and a matched real-data control identify settings where inter-class expansion provides benefits beyond adding repeated real samples.

2 Background

2.1 Identity-based shortcuts in EEG classification

EEG carries subject-specific patterns that can act as a shortcut for diagnostic prediction when each participant has one fixed label (Geirhos et al., 2020). Wang et al. (2024b) show this directly: permuting diagnostic labels across subjects while each subject's windows stay consistently labeled left EEGConformer's subject-dependent F1 for Parkinson's detection nearly unchanged (98.35% to 98.15%), so high performance did not require the intended disease–label relationship. The single-class-per-subject (SCPS) versus multi-class-per-subject (MCPS) distinction explains why the shortcut weakens when a subject contributes several classes (Li et al., 2026): in SCPS, subject identity determines the observed label, whereas in MCPS knowing the subject cannot resolve which class state produced a window, so the classifier must also use information that varies within that subject. Empirically, Brookshire et al. (2024) report accuracy dropping from 99.8% to 53.0% when moving from segment-based to subject-based evaluation for Alzheimer's detection, an SCPS task, against 79.1% to 65.1% for seizure detection, where a subject contributes both seizure and non-seizure windows; and Wang et al. (2024b) report EEGConformer's F1 falling by 0.90 percentage points for MCPS stimulus classification against 30.66 points for SCPS Parkinson's detection on unseen rather than seen subjects. SCPS label structure therefore adds an opportunity for identity-based overfitting beyond the distribution shifts that affect both task types, which motivates inter-class expansion, giving existing training subjects opposite-class examples so that recognizing the subject alone becomes less useful for predicting their labels. Whether that helps is the empirical question addressed here.

2.2 EEG generation for data augmentation

A common objective in EEG generation is more examples of a given class: Carrle et al. (2023) train a conditional Wasserstein GAN to synthesize EEG labeled as depression or healthy control, increasing sample counts and within-class variation, but a class label alone does not specify a subject identity, so more synthetic windows need not represent more distinct subjects. Torma & Szegletes (2025) generate motor-imagery and visual-evoked-potential signals, and Chen et al. (2024) reassemble transformer-diffusion segments with real EEG to build additional training examples; their reported sample diversity should be distinguished from explicit generation of new subject identities or of new diagnostic states for a fixed subject. Other methods tailor generation to an existing subject or domain, conditioning on a target subject's features (Fahimi et al., 2021), translating recordings toward a

target subject's distribution (Yin et al., 2023), or generating topographies for motor-imagery training in stroke patients (Xu et al., 2021), improving coverage of a target distribution rather than creating additional diagnostic classes for the same person. Closer to that objective, Pascual et al. (2019) generate patient-specific seizure EEG from non-seizure recordings, though seizure and non-seizure states are observed within patients in the source population, unlike the fixed-diagnosis structure of SCPS datasets; the follow-up EpilepsyGAN work also targets privacy preservation, so personalized generation need not preserve recognizable identity (Pascual et al., 2021). SPICE instead works on the diagnostic classes represented *within each training subject*: it aims to retain subject-specific characteristics while generating the opposite diagnostic class, weakening the exclusive subject–label association in SCPS data, and its contribution is that inter-class expansion objective rather than a larger sample count or a particular generative architecture.

3 METHOD

3.1 SUBJECT-PRESERVING INTER-CLASS EXPANSION

Let $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^{N}$ denote a training dataset, where $x_i \in \mathbb{R}^{C\times T}$ is an EEG window, $y_i \in \{0, 1\}$ is its diagnostic label, and s_i identifies its subject. In a single-class-per-subject (SCPS) dataset, all observations from subject s share one label y_s. This makes subject identity predictive of the label within the training set, even when the features identifying a person do not generalize to unseen subjects.

SPICE augments each eligible source window with a synthetic counterpart $(\widetilde{x}_i, 1 - y_i, s_i)$, changing diagnostic characteristics while retaining subject-specific information; arbitrary opposite-class examples would not serve this purpose, since the new examples must remain linked to an existing training subject. Subjects with accepted counterparts therefore acquire observations from both diagnostic classes, a construction we call *pseudo-MCPS*: the alternate-class observations are generated rather than recorded from the same person in another clinical state, which changes the training label support within subjects without assuming that all subject-related biases disappear. Because the dataset holds both diagnostic groups but no paired observations of the same person in both states, we formulate counterfactual generation as *unpaired translation*: SPICE specifies the desired translation and its acceptance criteria, and the CycleGAN implementation below provides one way to realize the framework.

3.2 UNPAIRED EEG TRANSLATION

We learn two directional generators G_0, G_1 with target-domain discriminators D_0, D_1 (Zhu et al., 2017), writing $\widetilde{x} = G_y(x)$ for a source window x of class y; each generator predicts a residual waveform and the discriminators score local temporal patches. The least-squares adversarial term $\mathcal{L}_{\text{adv}}$ asks that a translated window match the requested diagnostic domain, and cycle consistency $\mathcal{L}_{\text{cyc}}$ asks that translation in both directions reconstruct the source:

$$\mathcal{L}_{\text{adv}} = \sum_{y=0}^{1} \mathbb{E}_{x\sim\mathcal{D}_y}\left[\text{mean}\left[(D_{1-y}(G_y(x)) - 1)^2\right]\right], \tag{1}$$

$$\mathcal{L}_{\text{cyc}} = \sum_{y=0}^{1} \mathbb{E}_{x\sim\mathcal{D}_y}\left[\text{MAE}\left(G_{1-y}(G_y(x)), x\right)\right], \tag{2}$$

with MAE the elementwise mean absolute error and the adversarial mean taken over temporal patch scores. Neither objective alone explicitly requires the translated window to retain its source subject's characteristics. Appendix C.1 gives the generator and discriminator architectures in full.

3.3 PRESERVING SUBJECTS WHILE ENCOURAGING DIAGNOSTIC CHANGE

Subject preservation. We first train a temporal convolutional neural network (CNN) to identify training subjects and freeze its encoder E, whose 64-dimensional representation provides a reference for source–counterfactual similarity:

$$\mathcal{L}_{\text{subj}} = \sum_{y=0}^{1} \mathbb{E}_{x\sim\mathcal{D}_y}\left[1 - \cos\left(E(x), E(G_y(x))\right)\right]. \tag{3}$$

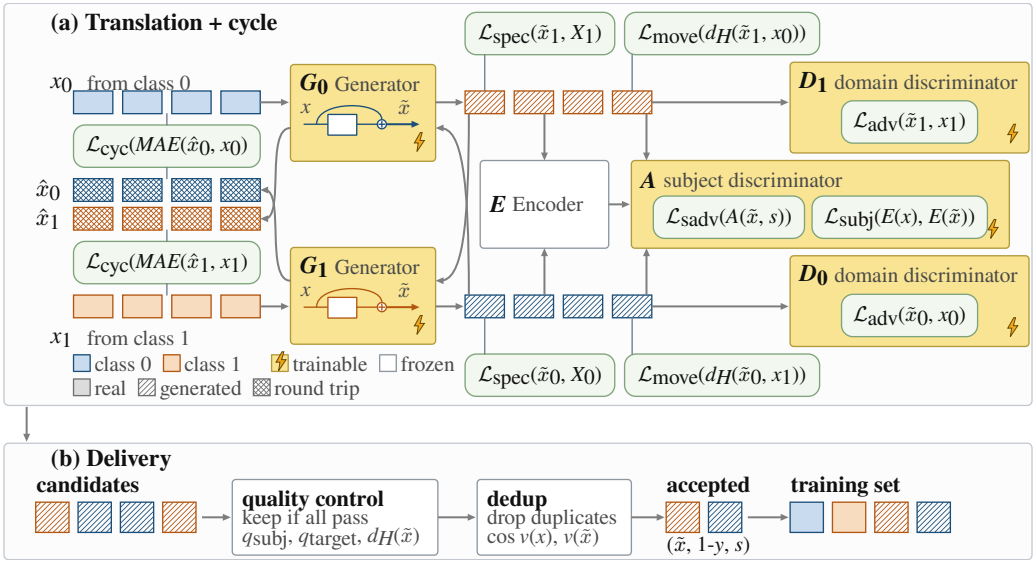


Figure 2: **The generation architecture.** (a) A real class-y window x_y is translated by the residual generator G_y into a counterpart $\widetilde{x}$ that the domain discriminator D_{1-y} scores adversarially, while the frozen encoder E and the trainable subject discriminator A tie it to its source subject; every objective sits where it is computed and each counterpart re-enters the other generator for the cycle return. (b) Screened candidates join the real rows in the training fold: the content that the matched real-data control isolates.

A trainable subject-conditioned discriminator $A(x, s)$ complements this fixed representation: it pools temporal convolutional features $h(x)$ and applies a scale and offset from a learned subject embedding, $A(x, s) = w^\top[(1 + \gamma_s) \odot h(x) + \beta_s] + b$, with real windows paired with their own subject identifiers and generated windows with their source identifiers. The generator minimizes $\mathcal{L}_{\text{sadv}} = \sum_{y=0}^{1} \mathbb{E}_{(x,s)\sim\mathcal{D}_y}[(A(G_y(x), s)-1)^2]$, which encourages compatibility with source-subject characteristics. Because subject and diagnosis are associated in the observed data, these learned constraints do not by themselves disentangle identity from disease.

Further generator terms. Two further terms shape the translated signal, defined in full in Appendix C.1: a spectral-matching term compares the window-level spectra of generated batches against unpaired real batches in the target domain, so that domain statistics are matched without pairing subjects across classes, and a displacement term penalizes insufficient waveform movement relative to the source and subtracts a moving average so that a constant offset cannot satisfy it. Neither term establishes diagnostic conversion on its own.

Combined objective. The implemented generator objective is

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + 2\mathcal{L}_{\text{cyc}} + 0.25\mathcal{L}_{\text{subj}} + \mathcal{L}_{\text{sadv}} + \mathcal{L}_{\text{spec}} + \mathcal{L}_{\text{move}}, \tag{4}$$

with generator, domain-discriminator and subject-discriminator updates alternating, and with frozen diagnostic scorers (below) assessing candidates after generation rather than supplying a classification gradient; Appendix C.1 records the discriminator objectives and the residual network in full. Figure 2 assembles both translation directions, the subject-side pair E and A, every objective and the delivery chain in one view.

3.4 Task-specific counterfactual selection

The goals of preserving subject information and changing diagnostic class provide a direct basis for screening candidates. For a source x of class y, we compute

$$q_{\text{subj}}(x, \widetilde{x}) = \cos(E(x), E(\widetilde{x})), \qquad q_{\text{target}}(\widetilde{x}) = \frac{1}{M}\sum_{j=1}^{M} Q_j(1 - y \mid \widetilde{x}), \tag{5}$$

where Q_j are two frozen diagnostic classifiers fitted on real training data and selected on real validation data, and we keep candidates with finite waveforms and diagnostic statistics, $q_{\text{subj}} > 0.7$,

$q_{\text{target}} > 0.5$ and $d_H(\widetilde{x}, x) \leq 0.9215$ to limit displacement from the source. Separately trained generators each propose candidates, and up to two are retained, starting with the smallest raw waveform MAE and then by the largest minimum H-space distance to those already retained (Appendix C.1 defines d_H); no minimum diversity is imposed.

Before downstream augmentation, the delivered rows are screened for near-copies in the fold's stored, training-only standardized space: a delivered row is kept only if the waveform cosine between it and its source window is below 0.9 (Appendix B). Near-copies remain among the delivered windows and their share is reported; the augmented training condition and its matched real-data control are both trained on the surviving rows alone. This screen discourages reproduction of the original waveform: retained candidates combine subject similarity with a changed signal and a predicted diagnosis. These scores are operational quality proxies, not proof of physiological counterfactual validity; the synthetic label remains the requested target $1 - y$, the source subject identifier is retained, and acceptance need not yield equal class counts or counterparts for every subject.

3.5 Training and downstream evaluation

All preprocessing statistics, auxiliary-network fitting, generator training and candidate generation are performed within each subject-disjoint fold, with each channel standardized on training-only statistics. Window geometry follows each cohort's own export (Table 2), so settings are not compared across cohorts; validation and test sets contain real recordings from subjects absent from that fold's training set. Generation stops on the best real-validation sample-level macro-F1 of a spectral multi-layer perceptron (MLP) trained on real training data plus the candidates generated so far (checkpoint selection through downstream utility rather than end-to-end classifier guidance). Appendix C.1 records the optimizer, the stopping rule, the per-fold budget and the selection thresholds in full.

The selected counterfactuals are then fixed for downstream comparisons. A separate classifier is trained on real-only data, on real data plus accepted counterfactuals, or on real data plus a matched real-data control that draws real training windows with replacement to match the generated rows row for row and class for class, so that only the content of the added rows differs. The number of generated rows a fold receives is capped at a fixed per-fold budget (Appendix C.1), and each training condition stops on its own patience, with classifier checkpoints selected by macro-F1 over the real validation windows before test evaluation. The primary outcome is patient-level fold-average macro-F1 (each test subject's window probabilities averaged, then each fold scored and the folds averaged), with sample-level performance reported beside it; every number names its aggregation, since pooled and fold-average values are not interchangeable.

4 Experiments

We evaluate SPICE to determine whether counterfactual augmentation enhances cross-subject disease detection and whether any observed benefit derives from counterfactual content rather than added sample exposure. As detailed in Section 4.1, each experiment compares models trained with generated counterfactuals against matched real-data controls receiving an identical volume of additional real data. Experiments span four clinical cohorts (ADFTD, ADHD-WMRI, CAUEEG, PD-RS), six model architectures, 23 settings, and 32 random seeds evaluated under strictly subject-disjoint five-fold cross-validation. Performance is measured primarily by fold-averaged patient-level macro-F1. Cohort details, implementation specifications, and the full tabular results are provided in Appendices A to F.

4.1 Experimental design

Three training conditions. Each cohort is split into five subject-disjoint folds, three for training, one for validation and one for test; every setting trains three classifiers on the same real training rows (Figure 3a). The first trains on those rows alone: the real-only baseline. The second adds generated windows, produced inside the fold by rewriting real windows of one class into windows of the other while keeping the subject identity of the source recording, with near-copies of their source windows removed before training (Figure 3b). The third adds the same number of extra real windows, drawn with replacement and matched in class composition: the matched real-data control (Figure 3c).

Metric and uncertainty. The reported quantity is the difference between the second condition and the third: comparing against the real-only baseline alone would confound the content of the generated windows with the fact of training on more windows, and the matched control removes

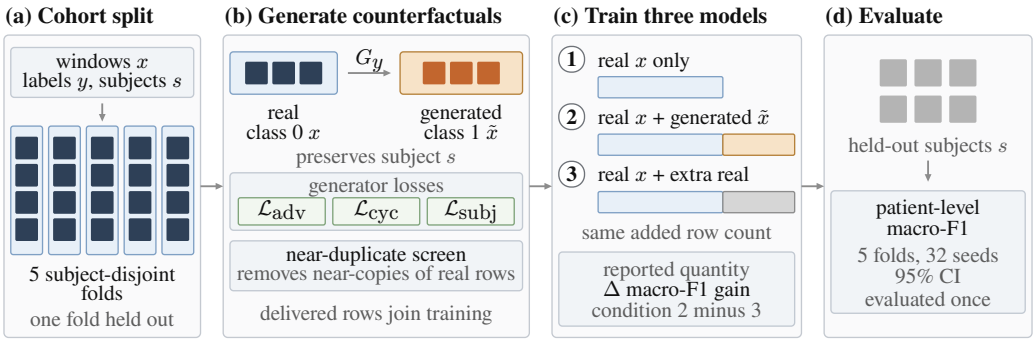


Figure 3: **The experimental design.** Subjects are divided into five subject-disjoint folds (a). Within each training fold a generator rewrites real windows of one class into windows of the other class while preserving subject identity, and the delivered rows pass a near-duplicate screen (b). Three classifiers train on the same real data, alone, with generated data, and with an equal amount of extra real data (c), and each is evaluated on subjects held out of its own fold (d).

that confound. The primary metric is patient-level macro-F1, computed per fold over subjects and averaged within a seed; we report the mean over 32 seeds with a 95% confidence interval, because the seed-to-seed spread dominates the uncertainty at these effect sizes, and we treat a difference of 2 points as the bar for a practical gain. The seeds, the comparison and the metric were fixed in a pre-specified analysis plan before the runs, and each setting's held-out fold is evaluated once, after model selection; each setting is evaluated on its own cohort and never compared across cohorts.

4.2 Main results

Table 1 gives the comparison for all 23 settings. Six settings improve clearly, with the whole interval above zero. On ADFTD the temporal convolutional network (TCN) gains $+3.27$ points over the matched real-data control (95% interval $[+1.90, +4.65]$) and EEGConformer $+3.23$ points ($[+1.90, +4.56]$); MedFormer gains $+23.18$ points ($[+19.71, +26.64]$), a training-stability result rather than a margin over a healthy baseline: both real-data-trained classifiers stall at chance here (Section 4.4). On ADHD-WMRI, CBraMod gains $+1.96$ points ($[+1.18, +2.73]$); on CAUEEG the TCN gains $+1.12$ points ($[+0.71, +1.53]$); on PD-RS the TCN gains $+1.06$ points ($[+0.12, +1.99]$). The TCN is the only architecture with a clear improvement on more than one cohort.

Seven further settings are positive on average, with intervals that leave the direction open, and three are negative with the whole interval below zero, ADHD-WMRI EEGConformer (-3.49), PD-RS CBraMod (-0.75) and CAUEEG LaBraM (-0.45), each discussed in Section 4.5; the remaining settings have intervals that include zero. Per-cohort results follow Table 1 (Figure 4), and Appendix D sets out every setting, cohort by cohort.

4.3 Content, not extra training

Generated windows change both what a classifier sees and how much it sees; the matched real-data control separates the two for the six clear settings (Table 7). The gain from the generated rows exceeds that from an equal number of extra real windows in all six, and in five the exposure contrast is small with an interval that includes zero; on PD-RS the extra real windows help as well ($+1.17$ points) and the generated windows still add $+1.06$ points beyond them. Utility $=$ content $+$ exposure holds exactly in the stored values, and both quantities are reported for every setting in Appendix D.

4.4 Ablations and robustness

Generated data work alongside real data. On the ADFTD TCN setting, a classifier trained on generated windows alone is 8.63 points below a real-data control of the same size ($[-10.43, -6.83]$) and 11.99 points below the real-only baseline ($[-14.06, -9.93]$): generated windows do not carry the task on their own, and their value in this study appears together with the real data (Appendix D).

Training stability on ADFTD MedFormer. On this setting both real-data-trained classifiers stall at chance (balanced accuracy 0.511 and 0.515, against 0.669 for the generated-data condition; its patient-level macro-F1 is 0.619). The mechanism behind the stall is outside this paper's scope;

Cohort	Architecture	Improvement (pt)	95% interval
ADFTD 65 subjects	TCN	$\mathbf{+3.2734}$	$[+1.8985, +4.6483]$
	Spectral-CNN	$+0.0311$	$[-0.9906, +1.0528]$
	MedFormer	$\mathbf{+23.1776}$	$[+19.7112, +26.6441]$
	EEGConformer	$\mathbf{+3.2276}$	$[+1.8963, +4.5590]$
	LaBraM	-1.2344	$[-2.5080, +0.0393]$
	CBraMod	$+0.3019$	$[-1.0891, +1.6930]$
ADHD-WMRI 59 subjects	TCN	$+1.4209$	$[-0.2604, +3.1021]$
	Spectral-CNN	-0.2651	$[-2.0258, +1.4956]$
	MedFormer	$+1.4789$	$[-0.3884, +3.3461]$
	EEGConformer	-3.4927	$[-5.9466, -1.0388]$
	LaBraM	-1.1687	$[-2.8657, +0.5282]$
	CBraMod	$\mathbf{+1.9550}$	$[+1.1849, +2.7251]$
CAUEEG 689 subjects (two-class)	TCN	$\mathbf{+1.12}$	$[+0.71, +1.53]$
	Spectral-CNN	-0.28	$[-0.82, +0.26]$
	EEGConformer	-0.63	$[-1.66, +0.39]$
	LaBraM	-0.45	$[-0.83, -0.08]$
	CBraMod	-0.06	$[-0.31, +0.19]$
PD-RS 149 subjects	TCN	$\mathbf{+1.06}$	$[+0.12, +1.99]$
	Spectral-CNN	$+0.6668$	$[-0.3128, +1.6463]$
	MedFormer	-0.11	$[-1.61, +1.40]$
	EEGConformer	$+0.7525$	$[-0.8784, +2.3834]$
	LaBraM	$+0.2920$	$[-1.2595, +1.8434]$
	CBraMod	-0.7481	$[-1.4736, -0.0227]$

Table 1: **Generated data against a matched amount of extra real data.** Improvement is the difference in patient-level macro-F1 (points) between training on real data plus generated data and on real data plus a matched amount of extra real windows, with 95% intervals over 32 seeds and five subject-disjoint folds per setting; bold marks settings whose interval lies entirely above zero, and values are printed at the precision of the analysis behind each row. Rows are read independently and are not ranked against one another; the CAUEEG MedFormer setting is still running and is not listed.

the generated windows kept the training run healthy as it ran, and the per-architecture counts of collapsed runs, together with this setting's earlier runs, are in Appendix D.

4.5 Limitations and future work

Three settings are negative. On ADHD-WMRI, EEGConformer loses 3.49 points ($[-5.95, -1.04]$), where the screen leaves the thinnest added content of the four cohorts and the test folds hold 11 to 12 subjects, so a few subjects move a fold's macro-F1. On PD-RS, CBraMod loses 0.75 points ($[-1.47, -0.02]$) with a utility contrast of essentially zero (-0.08); on CAUEEG, LaBraM loses 0.45 points ($[-0.83, -0.08]$) with every secondary metric agreeing in sign. Each loss has its own cause and next step: a more diverse generator and more subjects per test fold on ADHD-WMRI, a fine-tuned backbone on PD-RS, and a budget that scales with a cohort's real-data size on CAUEEG (set out, with the settings still open, in Appendix E).

5 Discussion

5.1 Interpretation and limitations

SPICE examines whether introducing diagnostic variation within training subjects can improve generalization beyond adding more examples of their observed class. The results are consistent with this strategy in several settings, with the TCN showing improvements over a matched real-data control on three cohorts. However, benefits vary across classifiers and datasets. Inter-class expansion is therefore a promising augmentation strategy whose usefulness depends on the generated examples and the model learning from them, rather than a general remedy for poor cross-subject generalization.

The matched real-data control helps distinguish the effect of generated examples from that of adding repeated real windows with the same class counts. It does not fully isolate this effect from optimization: each condition uses independent early stopping, so total training updates can differ. The large ADFTD MedFormer improvement warrants particular caution because both real-data baselines stall near chance. This result suggests an optimization benefit under the evaluated configuration, but does not establish superiority over a well-performing MedFormer baseline. Comparisons with matched update budgets and stable baselines would clarify these effects.

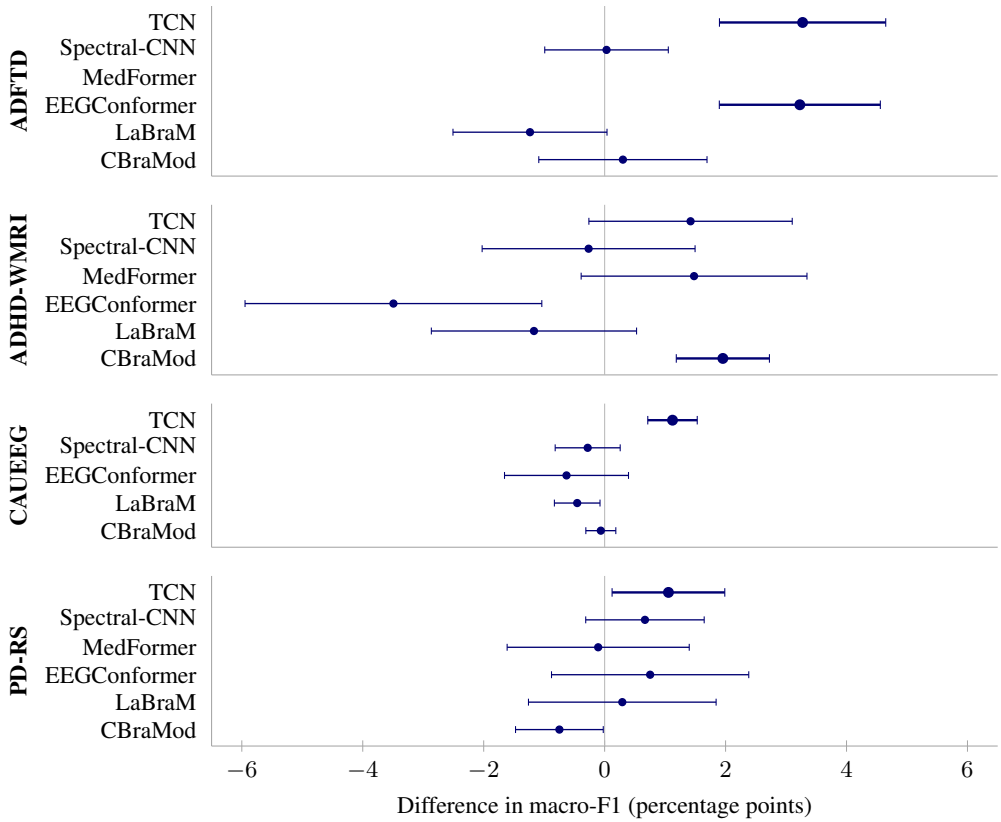


Figure 4: **Content contrast per architecture.** The quantity of Table 1 (difference in patient-level macro-F1, in points, between real data plus generated data and real data plus a matched amount of extra real data; 95% intervals over 32 seeds), one panel per cohort on a single shared x scale whose gray line marks zero; bolder rows are the six settings whose interval lies entirely above zero. The scale is cut at ± 6.5 points, so the $+23.18$-point ADFTD MedFormer row, whose interval reaches $+26.64$, lies beyond the cut and outside the plotted range; the estimate and interval of every other row lie inside the cut. That row recovers a run whose real-data baselines stall at chance, so it is a training-stability result, not a margin over a healthy baseline (Section 4.4).

Improved classification also does not, by itself, demonstrate reduced reliance on subject identity. SPICE changes the assigned label support within source subjects, while actual subject preservation and diagnostic conversion are assessed through learned proxies. In particular, the subject encoder used for selection also participates in generator training, and its representation may retain diagnostic information. Independent assessments of subject consistency and targeted ablations of the preservation objectives are needed to test the proposed mechanism. Without paired alternate-diagnosis recordings, the generated signals should be interpreted as constrained augmentation candidates rather than verified biological counterfactuals.

Candidate retention and augmentation size provide further context for the heterogeneous results. Near-copy filtering removes many candidates on ADHD-WMRI, whereas the fixed augmentation budget limits the fraction of generated data used on CAUEEG. These observations motivate studying candidate quality and augmentation dose, but do not establish why particular classifiers improve or deteriorate. Similarly, repeated training seeds characterize optimization variability on the existing subject splits; they do not resolve uncertainty about performance in new patient populations. Additional subject partitions and external-cohort evaluation would strengthen the assessment of generalization.

5.2 Future directions

The framework applies beyond the binary CycleGAN formulation evaluated here. For tasks involving several disorders or disease stages, a shared generator conditioned on the requested diagnostic class could replace separate pairwise translators. StarGAN provides an example of this

multi-domain design (Choi et al., 2018). An EEG adaptation would need to preserve source-subject characteristics while producing distinct, class-relevant changes across multiple target diagnoses.

Diffusion models offer another direction, following their image-generation results and emergence in time-series and EEG synthesis (Dhariwal & Nichol, 2021; Yuan & Qiao, 2024; Torma & Szegletes, 2025). Whereas each current generator maps a source window to one deterministic output, a source-conditioned diffusion model could sample multiple candidate counterfactuals. Subject-preservation constraints and target-class conditioning could guide these samples toward SPICE's objective, potentially increasing the useful diversity available after selection. Neither stochastic sampling nor a larger accepted pool guarantees better augmentation, however. These extensions should be assessed through independent generation-quality measures and downstream performance on unseen subjects, with controlled augmentation and training budgets.

6 Conclusion

We asked whether subject-preserving, label-inverting counterfactual windows add utility when a classifier trains alongside real data, and answered it against a matched real-data control across four cohorts. In settings spanning all four cohorts the generated content adds utility beyond an equal amount of extra real data, while others remain open; the generated rows do not carry the task on their own, and the size of a gain follows how much generated content reaches training. The same configuration and evaluation protocol apply to further cohorts; a more diverse generator, more seeds per setting and wider architecture coverage would settle the settings still open.

AI use statement

In this work, generative AI tools assisted in writing and editing the manuscript and in developing and running the experiment code, but no reported result depends on AI-generated output: the analysis plan, the comparisons and the metrics were fixed before the runs, and every reported number is recomputed by code from saved per-window predictions; no other tasks with required disclosure apply. We reviewed all AI-assisted work: the code was tested, every reported value re-derived from the saved predictions, and the text checked against the tables and figures; we take responsibility for the final content, including text, claims or artifacts produced with the aid of generative AI. See the ICLR 2027 AI Policy for Authors for more details.

Ethics statement

This work uses only publicly available, de-identified human EEG datasets (Appendix A), used under their original terms; no new data were collected, no participants were recruited, and no new ethical approval was required. Because the framework generates synthetic EEG windows that retain the identity characteristics of their source recordings, and identity-linked generation can be misused, we describe the processing and near-duplicate screening in full and will release the generated data only together with the code and analysis plans. Our claims concern relative classifier utility under subject-disjoint evaluation, not clinical or diagnostic use.

Reproducibility statement

Every number reported here comes from the protocol described in Sections 3 and 4, and each is backed by artifacts to be released with the paper: each cohort is split into five subject-disjoint folds; all preprocessing statistics, generator fitting and candidate selection use training rows only; every model and checkpoint is selected by macro-F1 over the real validation windows; and each setting's held-out fold is opened once, after its models are finished. Each setting trains 32 seeds per training condition, and the release will include the per-seed test probabilities and labels from which each metric is recomputed. The comparison, the seeds and the metric were fixed in a pre-specified analysis plan before any model was trained, and the checks behind the numbers are part of the analysis code included in the release. The four cohorts are public datasets used under their original terms and are available from their original repositories, which distribute the recordings rather than this paper; the generated data, code and analysis plans will be released with the paper.

REFERENCES

Caroline L Alves, Aruane M Pineda, Kirstin Roster, Christiane Thielemann, and Francisco A Rodrigues. Eeg functional connectivity and deep learning for automatic diagnosis of brain disorders: Alzheimer's disease and schizophrenia. *Journal of Physics: Complexity*, 3(2):025001, 2022.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Carolin Breitling-Ziegler, Jana Tegelbeckers, Hans-Henning Flechtner, and Kerstin Krauel. Economical assessment of working memory and response inhibition in adhd using a combined n-back/nogo paradigm: An erp study. *Frontiers in Human Neuroscience*, 14:322, 2020.

Geoffrey Brookshire, Jake Kasper, Nicholas M Blauch, Yunan Charles Wu, Ryan Glatt, David A Merrill, Spencer Gerrol, Keith J Yoder, Colin Quirk, and Ché Lucero. Data leakage in deep learning studies of translational eeg. *Frontiers in Neuroscience*, 18:1373515, 2024.

Friedrich Philipp Carrle, Yasmin Hollenbenders, and Alexandra Reichenbach. Generation of synthetic eeg data for training algorithms supporting the diagnosis of major depressive disorder. *Frontiers in Neuroscience*, 17:1219133, 2023.

James F. Cavanagh and Darin Brown. EEG: reinforcement learning in parkinson's. OpenNeuro dataset ds003506, 2021.

Mingzhi Chen, Yiyu Gui, Yuqi Su, Yuesheng Zhu, Guibo Luo, and Yuchao Yang. Improving eeg classification through randomly reassembling original and generated data with transformer-based diffusion models. *arXiv preprint arXiv:2407.20253*, 2024.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797. IEEE, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Fatemeh Fahimi, Strahinja Dosen, Kai Keng Ang, Natalie Mrachacz-Kersting, and Cuntai Guan. Generative adversarial networks-based data augmentation for brain–computer interface. *IEEE Transactions on Neural Networks and Learning Systems*, 32(9):4039–4051, 2021. doi: 10.1109/TNNLS.2020.3016666.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, pp. 16405–16426, 2024.

Min-jae Kim, Young Chul Youn, and Joonki Paik. Deep learning-based eeg analysis to classify normal, mild cognitive impairment, and dementia: Algorithms and dataset. *NeuroImage*, 272: 120054, 2023.

Taida Li, Yujun Yan, Fei Dou, Wenzhan Song, and Xiang Zhang. Cross-subject generalization for eeg decoding: a survey of deep learning methods. *Progress in Biomedical Engineering*, 8(2): 022013, 2026.

Alfredo Lucas, Andrew Revell, and Kathryn A Davis. Artificial intelligence in epilepsy—applications and pathways to the clinic. *Nature Reviews Neurology*, 20(6):319–336, 2024.

Andreas Miltiadous, Katerina D Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G Tsalikakis, Pantelis Angelidis, Markos G Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, et al. A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg. *Data*, 8(6):95, 2023.

Damian Pascual, Amir Aminifar, David Atienza, Philippe Ryvlin, and Roger Wattenhofer. Synthetic epileptic brain activities using generative adversarial networks. *arXiv preprint arXiv:1907.10518*, 2019.

Damian Pascual, Alireza Amirshahi, Amir Aminifar, David Atienza, Philippe Ryvlin, and Roger Wattenhofer. Epilepsygan: Synthetic epileptic brain activities with privacy preservation. *IEEE Transactions on Biomedical Engineering*, 68(8):2435–2446, 2021.

Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of Neural Engineering*, 16(5):051001, 2019.

Arun Singh, Rachel Cole, Arturo Espinoza, Jim Cavanagh, and Nandakumar Narayanan. Rest eyes open. OpenNeuro dataset ds004584, 2023.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023.

Szabolcs Torma and Luca Szegletes. Generative modeling and augmentation of eeg signals using improved diffusion probabilistic models. *Journal of Neural Engineering*, 22(1):016001, 2025.

Hanneke Van Dijk, Guido Van Wingen, Damiaan Denys, Sebastian Olbrich, Rosalinde Van Ruth, and Martijn Arns. The two decades brainclinics research archive for insights in neurophysiology (tdbrain) database. *Scientific Data*, 9(1):333, 2022.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International Conference on Learning Representations*, pp. 75310–75346, 2025a.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36314–36341, 2024a.

Yihe Wang, Taida Li, Yujun Yan, Wenzhan Song, and Xiang Zhang. What causes performance degradation in cross-subject eeg classification? *arXiv preprint arXiv:2410.03057*, 2024b.

Yihe Wang, Nan Huang, Nadia Mammone, Marco Cecchi, and Xiang Zhang. LEAD: an EEG foundation model for Alzheimer's disease detection. *arXiv preprint arXiv:2502.01678*, 2025b.

Fangzhou Xu, Fenqi Rong, Jiancai Leng, Tao Sun, Yang Zhang, Siddharth Siddharth, and Tzyy-Ping Jung. Classification of left-versus right-hand motor imagery in stroke patients using supplementary data generated by cyclegan. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2417–2424, 2021.

Kang Yin, Byeong-Hoo Lee, Byoung-Hee Kwon, and Jeong-Hyun Cho. Target-centered subject transfer framework for eeg data augmentation. In *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*, pp. 1–4. IEEE, 2023. doi: 10.1109/BCI57258.2023.10078578.

Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b5b66077d016c037576cc56a82f97f66-Abstract-Conference.html.

Marzieh Zare. A negative-control protocol for clinical eeg foundation-model benchmarks: Dataset identity and external-cohort stress testing, 2026. URL https://arxiv.org/abs/2607.24519.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251. IEEE, 2017.

Cohort	Subjects	Windows	Channels	Timestamps	Rate and intake
ADFTD	65	53,150	19	400	200 Hz native
CAUEEG	689 (two-class)	475,852	19	400	200 Hz native, no resampling
PD-RS	149	23,690	60	400	200 Hz; a rate filter kept 23,690 of 41,174 exported rows
ADHD-WMRI	59 (25 control, 34 ADHD)	13,540	21	200	200 Hz native; Combined-Task recording
RLPD	56	21,510	59	200	200 Hz native; single-task recording (task 3)
TDBrain	204	28,440	26	400	200 Hz native; merged resting-task recordings
APAVA	23	3,315	19	200	128 Hz resampled once at intake to 200 Hz
ADSZ	24 (48 rec.)	408	19	200	same intake as APAVA

Table 2: **Cohorts, intake counts and window geometry.** Window geometry is each cohort's own export. APAVA and ADSZ contribute intake counts only; no results are reported for them in this paper. Cohort sources: ADFTD (Miltiadous et al., 2023); CAUEEG (Kim et al., 2023); PD-RS (Singh et al., 2023); ADHD-WMRI (Breitling-Ziegler et al., 2020); RLPD (Cavanagh & Brown, 2021); TDBrain (Van Dijk et al., 2022); APAVA and the LEAD-family intake (Wang et al., 2025b); ADSZ (Alves et al., 2022).

A Datasets and cohorts

This paper reports results on four cohorts (ADFTD, ADHD-WMRI, CAUEEG and PD-RS) and lists the remaining cohorts that entered the same processing path, with their intake counts. Table 2 gives the intake counts and the window geometry of every cohort. Window geometry differs by cohort rather than by experimental choice: each shape is the cohort's own export.

ADFTD (Miltiadous et al., 2023) is the two-class Alzheimer's-disease against healthy-control cohort drawn from a release that also carries frontotemporal-dementia recordings: 65 subjects and 53,150 windows of 19 channels by 400 timestamps at 200 Hz, with the five folds reusing the cohort's established subject-disjoint partition.

CAUEEG (Kim et al., 2023) has 689 subjects in the two-class subset used here (459 and 230) and 475,852 windows of 19 channels by 400 timestamps at 200 Hz, used without resampling.

PD-RS (Singh et al., 2023) has 149 subjects and 23,690 windows of 60 channels by 400 timestamps at 200 Hz after a rate filter that dropped 11,729 windows recorded at 100 Hz and 5,755 at 50 Hz, and no window was dropped by label; 8,253 windows carry class 0 and 15,437 carry class 1.

ADHD-WMRI (Breitling-Ziegler et al., 2020) has 59 subjects (25 control and 34 ADHD) and 13,540 windows of 21 channels by 200 timestamps at 200 Hz, on the CombinedTask recording.

RLPD (Cavanagh & Brown, 2021) has 56 subjects and 21,510 windows of 59 channels by 200 timestamps at 200 Hz, on the single-task recording (task 3), and **TDBrain** (Van Dijk et al., 2022) has 204 subjects and 28,440 windows of 26 channels by 400 timestamps at 200 Hz, on the merged resting-task recordings. **APAVA** (Wang et al., 2025b) and **ADSZ** (Alves et al., 2022) contribute intake counts only: 23 subjects with 3,315 windows and 24 subjects with 48 recordings (408 windows), 19 channels by 200 timestamps, recorded at 128 Hz and resampled once at intake to 200 Hz. No results are reported for these four cohorts here.

B Data generation details

The construction, schematically. Figure 1 shows the SCPS-to-pseudo-MCPS construction behind the generator.

Within each fold, a generator fitted on that fold's training rows only rewrites real windows of one class into windows of the other class, and the generated windows keep the subject identity of the recording they came from (Section 3). The delivered windows are then screened: a generated window that is a near-copy of its source window is removed before training, and the removed windows are claimed by no result in this paper. The screen tests waveform similarity in the fold's stored, training-only standardized space: with $v(x)$ flattening a window across channels and time and subtracting its overall mean, a delivered row is kept only if $\cos(v(x), v(\widetilde{x})) < 0.9$ against its source

Cohort	Generated windows	Removed as near-duplicates	Windows added per fold, after the screen and the per-fold budget
ADFTD	92,178	0.526 (all folds)	9,780 / 9,867 / 8,971 / 6,056 / 9,064
PD-RS	46,517	0.074 (fold 1: 0.207)	7,279 / 8,072 / 8,971 / 6,056 / 6,687; limited by supply on folds 1, 2 and 5
CAUEEG	641,923	0.460 (all folds)	9,780 / 9,867 / 8,971 / 6,056 / 9,064; the budget binds on all five folds (346,873 windows survive the screen)
ADHD-WMRI	26,014	0.577 (all folds; fold 3: 0.983)	2,997 / 1,700 / 127 / 3,798 / 2,390; fold 3 adds 127 windows
RLPD	47,249	0.090 (folds 1 to 3 above the line)	6,905 / 7,151 / 5,652 / 6,056 / 9,064
TDBrain	19,456	0.634 (all folds)	1,531 / 957 / 1,294 / 2,206 / 1,140

Table 3: **The generated data, per cohort.** One cohort per row; rows are read independently and are not ranked against one another. "Removed as near-duplicates" is the share of generated windows that the screen removes before training because they are near-copies of their source windows; the removed windows are claimed by no result in this paper. "Windows added per fold" is what reaches training after the screen and the per-fold budget: a fold adds the budget where enough windows survive, and only what survives where the supply is smaller. RLPD's fold-level near-duplicate shares are 0.137, 0.145 and 0.103 against a 0.090 cohort total, and its folds 4 and 5 are budget-bound; ADFTD's five folds carry the same per-fold budget, which is why its added counts match CAUEEG's; TDBrain has no budget-bound fold.

window. Table 3 gives the delivered counts, the removed shares, and the number of windows each fold adds after the screen and the per-fold budget.

The added data are small next to the real training rows in every cohort, and how small differs by cohort: TDBrain adds 7,128 windows to 85,320 real training windows, ADHD-WMRI 11,012 to 40,620, and RLPD 34,828 to 64,530. Where a fold's budget exceeds what survives the screen, the fold adds only what survives (PD-RS folds 1, 2 and 5, and ADHD-WMRI fold 3, which adds 127 windows); where the surviving supply exceeds the budget, the budget binds (all five CAUEEG folds, where 346,873 windows survive the screen and 43,738 are added).

C Architectures and implementation

Table 4 lists the six architectures and their configurations. Four rules hold the comparison fixed. All training conditions within a setting see the same real rows and labels, so the augmentation is the only difference between them. All settings of one cohort see the same rows, so the architecture is the only difference between them. Each architecture's preprocessing is fitted on training rows only, wherever a statistic is needed. Every setting also trains with the same objective and is evaluated the same way: plain cross-entropy with no auxiliary term, no class weighting and no label smoothing, the same primary metric, and the same rule for selecting the checkpoint (best macro-F1 over the real validation windows), with the held-out fold evaluated once afterwards. Training batches are 128 in every setting except CBraMod's 64, which follows its published linear-probe configuration.

Two properties of the setup qualify the numbers. CBraMod's input pipeline applies a declared microvolt rescale and is therefore not the other five architectures' pipeline. External evidence on the pretrained foundation models is negative: on CAUEEG, classical features reach 0.734 macro-AUROC (area under the receiver operating characteristic curve, AUROC) against CBraMod's 0.669 (Zare, 2026); this negative external evidence applies to the CBraMod settings and is not extended to LaBraM.

C.1 Generator and selection details

The residual generator. Each generator predicts $\widetilde{x} = G_y(x) = x + \Delta_{\theta_y}(x)$, a residual waveform from a length-preserving temporal convolutional network with 32 hidden channels and five blocks using dilations $1, 2, 4, 8, 16$; each block holds two kernel-five convolutions with channel-wise layer normalization and SiLU activations, and a convolutional stem and output head map between EEG channels and hidden features. Initializing the output head near zero starts training near the identity mapping. The source waveform is the sole generator input, the translation direction is determined by which network is used, and each network is deterministic.

Architecture	Configuration
TCN	width 32, depth 5, embedding 64; its configuration follows an earlier run rather than being re-tuned
Spectral-CNN	width 64, depth 1, bins 32; in-house classifier
MedFormer	width 128, depth 6, 8 heads, d_{ff} 256, patch lengths $[2, 4, 8]$, dropout 0.1, GELU activation
EEGConformer	width 40, depth 6, 10 heads, kernel 20, max-pool 60, stride 12, dropout 0.5
LaBraM	pretrained checkpoint, drop path 0.1, channel tokens indexed by the fixed 19-channel order
CBraMod	pretrained checkpoint, linear probe, batch 64, learning rate 5×10^{-4}, backbone rate 1×10^{-4}, AdamW, weight decay 0.05, cosine schedule over 50 epochs, clipping 1.0

Table 4: **Architecture register, one architecture per row.** Five of the six see the fold's train-only standardized space; CBraMod's path applies a declared microvolt rescale, so its input pipeline is not the other five architectures' and no statement may imply otherwise. Architecture sources: TCN (Bai et al., 2018); Spectral-CNN has no external source; MedFormer (Wang et al., 2024a); EEGConformer (Song et al., 2023); LaBraM (Jiang et al., 2024); CBraMod (Wang et al., 2025a).

The discriminators. The domain discriminators use five strided temporal convolutional layers and output local real/fake scores; each is updated with the least-squares loss $\frac{1}{2}\{\mathrm{mean}[(r-1)^2] + \mathrm{mean}[f^2]\}$ on its real and detached generated scores, summed across both directions for the subject discriminator. The subject encoder is a temporal CNN trained to identify training subjects, then frozen; its 64-dimensional representation is used only through cosine similarity, and the encoder itself does not explicitly normalize its output. Its parameters remain fixed, but gradients through the generated embedding update the generator. The subject-conditioned discriminator uses γ_s, β_s from a learned subject embedding per subject and a scalar weight w and bias b; real windows are paired with their own subject identifiers and generated windows with their source identifiers. Generator, domain-discriminator and subject-discriminator updates alternate, and the generator objective weights the cycle term at 2 and the subject-similarity term at 0.25.

Stopping rule and budget. From epoch 10 on, a spectral MLP is retrained every five epochs on real training data plus the candidates generated so far, and its best real-validation sample-level macro-F1 decides when generation stops: after four checks without improvement exceeding 10^{-4}. Candidates are dropped unless their waveforms and diagnostic statistics are finite, $q_{\text{subj}} > 0.7$, $q_{\text{target}} > 0.5$ and $d_H(\widetilde{x}, x) \leq 0.9215$; the displacement definition uses $W = \mathrm{round}(1.5 f_s)$ and replicate padding in the moving average. The number of generated rows a fold receives is capped at a fixed per-fold budget, highest-quality rows first and without padding, so that the number of added rows is a stated design choice rather than an artifact of generator supply, and each training condition stops on its own patience. The generators and discriminators are optimized with Adam at learning rate 2×10^{-4} and batch size 128 per diagnostic domain. The optimizer-update counts of each training condition are reported beside every contrast.

The spectral and displacement terms, in full. With $P(x) = \log(|\,\mathrm{rFFT}_{\mathrm{ortho}}(x)|^2 + \epsilon)$ retaining frequencies from 0.5 to 45 Hz, and with batch means and population standard deviations μ_P, σ_P at each channel and frequency, the spectral term is $\mathcal{L}_{\text{spec}} = S(G_0(X_0), X_1) + S(G_1(X_1), X_0)$ for unpaired batches X_0 and X_1, with $S(U, V) = \mathrm{MAE}(\mu_P(U), \mu_P(V)) + 0.1\,\mathrm{MAE}(\sigma_P(U), \sigma_P(V))$ numerically stabilized and MAE the elementwise mean absolute error; distributions are compared rather than an arbitrary pairing between subjects in different classes. With $H(\delta) = \delta - \mathrm{MovingAverage}_W(\delta)$, $d_H(u, v) = \mathrm{mean}_{c,t}\,|H(u - v)|$ and $W = \mathrm{round}(1.5 f_s)$ under replicate padding, the displacement term is $\mathcal{L}_{\text{move}} = \sum_{y=0}^{1} \mathbb{E}_{x \sim \mathcal{D}_y}[\max\{0, m - d_H(G_y(x), x)\}]$ with $m = 0.28$ in standardized waveform units; subtracting the moving average prevents a constant offset from satisfying the objective, though it is not a strict band-pass filter. The term exists because subject preservation and cycle reconstruction can favor small changes, whereas useful opposite-class augmentation requires more than copying the source with a new label, so insufficient waveform movement relative to the source is penalized directly.

D Full per-setting results

Table 5 repeats the main results with the seed-level detail behind them: the number of seeds whose difference is positive, negative or exactly zero, the number of folds whose mean is positive, and the

smallest difference the design can detect with 80% power at the printed precision. Table 6 gives the per-fold means, Table 7 the utility and exposure contrasts together with the training lengths, and Table 8 where a secondary metric disagrees with the primary. Utility = content + exposure holds exactly in the stored values; entries may not sum at the printed precision.

Training lengths. The three training conditions stop on their own validation patience, so their training lengths are not matched in any setting. The ratios are reported in Table 7, and differences in training length are never taken as content conclusions. On the CAUEEG TCN setting the per-condition totals are 2,432,196 (real-only), 2,886,974 (generated) and 2,296,693 (matched real) updates, with a per-seed generated-to-real ratio ranging from 0.8956 to 1.7826.

The counterfactual-only ablation, in full. On the ADFTD TCN setting, the classifier trained on generated windows alone trains on 27% of the rows the real-only baseline uses ($43{,}738$ against $159{,}450$, summed over the five folds), and 24 of its 32 seeds carry the analysis plan's degenerate-cell flag and stay in every average. Both controls refer to the same fold structure as the augmented conditions.

Per-cohort results, in full. ADFTD (65 subjects, 19 channels, 53,150 windows): the TCN ($+3.27$ points), EEGConformer ($+3.23$) and MedFormer ($+23.18$, the training-stability result) improve clearly; Spectral-CNN ($+0.03$) and CBraMod ($+0.30$) sit close to zero with intervals that include zero, and LaBraM is negative on average (-1.23) with an interval that just includes zero. ADHD-WMRI (59 subjects, 21 channels, 13,540 windows): CBraMod improves clearly ($+1.96$ points, $[+1.18, +2.73]$); the TCN ($+1.42$) and MedFormer ($+1.48$) are positive on average with intervals that include zero; Spectral-CNN (-0.27) sits close to zero, and LaBraM is negative on average (-1.17); EEGConformer is negative, and the screen removes most of this cohort's generated windows (0.577 overall and 0.983 in fold 3, which adds 127 windows), so the added content is the thinnest of the four by row count. CAUEEG (689 subjects, 19 channels, 475,852 windows): the TCN improves clearly ($+1.12$ points, $[+0.71, +1.53]$) and is this cohort's one clear positive; Spectral-CNN (-0.28), EEGConformer (-0.63) and CBraMod (-0.06) sit close to zero; LaBraM is negative (-0.45); the per-fold budget rather than the supply limits the added data on all five folds (346,873 generated windows survive the screen; 43,738 are added), and this is the study's largest cohort by window count. PD-RS (149 subjects, 60 channels, 23,690 windows): the TCN improves clearly ($+1.06$ points, $[+0.12, +1.99]$); Spectral-CNN ($+0.67$), EEGConformer ($+0.75$) and LaBraM ($+0.29$) are positive on average with intervals that include zero; MedFormer sits close to zero (-0.11); CBraMod is negative; extra real windows help on this cohort as well ($+1.17$ points at matched exposure), and the generated windows add $+1.06$ points on top of them. These are the results summarized in Section 4.2.

Two rows come from replication runs. The CAUEEG TCN row and the PD-RS TCN row come from replication runs; both rows are printed at the precision of their own analyses, and no other row is compared with them.

Fold-level properties. The properties below travel with the per-cohort results of Section 4.2. ADFTD fold 4 carries the cohort's class imbalance (653 windows of one class against 5,403 of the other, a positive-class share of 0.892). ADHD-WMRI fold 3 adds 127 generated windows after the screen, so its column of Table 6 rests on little added content. ADHD-WMRI test folds hold 11 or 12 subjects, so a single one-sided run can dominate a fold: the numbers of each setting's 480 training runs (32 seeds on five folds under three training conditions) in which at least 90% of the fold's test subjects are assigned the same predicted class are 112 for EEGConformer, 88 for CBraMod, 86 for TCN, 48 for MedFormer, 34 for LaBraM and 22 for Spectral-CNN, kept as counted and never weighted.

Training stability on ADFTD MedFormer, in full. In 334 of the setting's 480 training runs at least 90% of the fold's test subjects are assigned the same predicted class (in 317 of them, every test subject is), the largest such count on ADFTD; the counts for the cohort's five other architectures are EEGConformer 74, TCN 5, LaBraM 1, CBraMod 0 and Spectral-CNN 0. The same behaviour appeared in this setting's earlier runs, where both real-data-trained runs held their training loss at the constant-prediction level for the whole run while validation macro-F1 sat at the majority-class level (0.3399). Both are kept as counted, and the setting's patient-level balanced-accuracy and macro-F1 averages are in Section 4.4.

The narrow interval on ADHD-WMRI CBraMod. Its interval is narrow because folds 2 and 4 are exactly zero and the cohort's test folds hold 11 to 12 subjects, so the per-subject scores take few distinct values across seeds; its AUROC score carries the opposite sign (Table 8).

Cohort	Architecture	Improvement ± sd (pt)	95% interval	Seeds +/−/0	Folds > 0	Detectable (pt)
ADFTD	TCN	$+3.2734 \pm 3.8134$	$[+1.8985, +4.6483]$	26/6/0	4/5	1.8876
	Spectral-CNN	$+0.0311 \pm 2.8339$	$[-0.9906, +1.0528]$	14/14/4	3/5	1.4027
	MedFormer	$+23.1776 \pm 9.6147$	$[+19.7112, +26.6441]$	32/0/0	5/5	4.7590
	EEGConformer	$+3.2276 \pm 3.6926$	$[+1.8963, +4.5590]$	27/5/0	3/5	1.8278
	LaBraM	-1.2344 ± 3.5326	$[-2.5080, +0.0393]$	12/20/0	2/5	1.7485
	CBraMod	$+0.3019 \pm 3.8582$	$[-1.0891, +1.6930]$	20/11/1	2/5	1.9097
ADHD-WMRI	TCN	$+1.4209 \pm 4.6632$	$[-0.2604, +3.1021]$	23/8/1	3/5	2.3082
	Spectral-CNN	-0.2651 ± 4.8835	$[-2.0258, +1.4956]$	15/17/0	2/5	2.4172
	MedFormer	$+1.4789 \pm 5.1790$	$[-0.3884, +3.3461]$	21/11/0	3/5	2.5635
	EEGConformer	-3.4927 ± 6.8063	$[-5.9466, -1.0388]$	8/23/1	0/5	3.3689
	LaBraM	-1.1687 ± 4.7067	$[-2.8657, +0.5282]$	13/18/1	2/5	2.3297
	CBraMod	$+1.9550 \pm 2.1359$	$[+1.1849, +2.7251]$	21/4/7	2/5	1.0572
CAUEEG	TCN	$+1.12 \pm 1.13$	$[+0.71, +1.53]$	28/4/0	4/5	0.5614
	Spectral-CNN	-0.28 ± 1.49	$[-0.82, +0.26]$	12/20/0	1/5	0.74
	EEGConformer	-0.63 ± 2.84	$[-1.66, +0.39]$	12/20/0	1/5	1.41
	LaBraM	-0.45 ± 1.05	$[-0.83, -0.08]$	10/22/0	1/5	0.52
	CBraMod	-0.06 ± 0.69	$[-0.31, +0.19]$	15/17/0	3/5	0.34
PD-RS	TCN	$+1.06 \pm 2.58$	$[+0.12, +1.99]$	20/12/0	5/5	1.28
	Spectral-CNN	$+0.6668 \pm 2.7169$	$[-0.3128, +1.6463]$	18/14/0	3/5	1.3448
	MedFormer	-0.11 ± 4.18	$[-1.61, +1.40]$	18/14/0	2/5	2.0671
	EEGConformer	$+0.7525 \pm 4.5234$	$[-0.8784, +2.3834]$	20/12/0	3/5	2.2390
	LaBraM	$+0.2920 \pm 4.3032$	$[-1.2595, +1.8434]$	16/16/0	2/5	2.1300
	CBraMod	-0.7481 ± 2.0121	$[-1.4736, -0.0227]$	9/23/0	2/5	0.9959

Table 5: **The content contrast with its seed-level detail.** Improvement is the patient-level macro-F1 difference, in percentage points, of the generated-data condition against the matched real-data control; each row rests on 32 seeds and five subject-disjoint folds, and every setting has its full seed set. "Seeds +/−/0" counts the per-seed differences that are positive, negative or exactly zero (zeros on neither side), and "Folds > 0" counts folds whose mean is positive. "Detectable" is the smallest difference this design can detect with 80% power at the printed precision, $2.8\,\mathrm{sd}/\sqrt{n}$ from the setting's own measured standard deviation. Values are printed at the precision of the analysis behind each row. The CAUEEG MedFormer setting is still running and is therefore absent from this table. Rows are read independently and are not ranked against one another.

E Reproducibility and compute

The three negative settings, in full. On ADHD-WMRI, EEGConformer loses 3.49 points: the screen removes most of the generated windows there, so the added content was thinnest exactly where the loss appears, and its test folds hold 11 to 12 subjects, so a few subjects move a fold's macro-F1. On PD-RS, CBraMod loses 0.75 points while its utility contrast is essentially zero (-0.08): a frozen backbone with a linear probe can only move at the read-out layer. On CAUEEG, LaBraM loses 0.45 points, with every secondary metric agreeing in sign, where the same per-fold budget adds the smallest share of new data. The next steps for each setting are set out in Section 4.5.

The settings still open, and their next steps. Seven settings are positive on average with intervals that include zero and need more seeds before their direction is settled (Tables 5 and 6); on ADFTD a fold with a more even class mix would help, and fold 4's imbalance is reported above. Per-fold means change sign inside several settings, and no fold is weighted or trimmed. The training lengths of the three conditions are not matched in any setting, and the length ratios are in Table 7; a length-matched variant is worth testing. Coverage is incomplete by design: four cohorts and 23 settings are analyzed here, two further cohorts with generated data (RLPD and TDBrain) have no results yet, and more architectures per cohort are warranted.

F Raw metric values

The four tables below print the absolute values behind the contrasts of Table 1, one table per cohort, with one sub-row per training condition: the real-only baseline (RO), the generated-data condition (CF) and the matched real-data control (CT). Every metric is recomputed from the saved per-window test predictions, and every entry is at patient level: a mean over the setting's 32 seeds together with its sample standard deviation (one degree of freedom), each seed first averaged over the five subject-disjoint folds. macro-F1 is the primary metric and the four remaining columns (accuracy, balanced accuracy, AUROC and positive-class F_1) are the secondary metrics the saved predictions carry. Means and standard deviations are printed to four decimals, so the difference of the generated and matched-real printed means recovers the corresponding contrast of Table 1 to at most a hundredth

Cohort	Architecture	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
ADFTD	TCN	+6.0461	+5.7869	+4.5385	+2.0253	−2.0298
	Spectral-CNN	+1.5370	−0.0353	+0.2456	−1.9920	+0.4003
	MedFormer	+27.5859	+39.4411	+25.7929	+22.4057	+0.6626
	EEGConformer	+4.4125	−0.4412	+10.4288	−3.4726	+5.2107
	LaBraM	+3.1730	−6.9182	+3.1494	−5.3081	−0.2679
	CBraMod	−3.2527	−4.4725	−1.0211	+1.8823	+8.3737
ADHD-WMRI	TCN	+1.9752	−3.7859	−2.1964	+4.0297	+7.0816
	Spectral-CNN	−0.3263	+5.2856	+0.8296	−1.8252	−5.2894
	MedFormer	−1.7616	+2.2219	+0.0899	−0.7245	+7.5686
	EEGConformer	−0.6674	−4.3078	−1.9849	−1.4819	−9.0216
	LaBraM	−2.6611	+0.3361	−0.9767	−4.3210	+1.7789
	CBraMod	+0.4369	0.0000	−1.3175	0.0000	+10.6556
CAUEEG	TCN	+1.3632	+0.9832	+2.9075	+2.4980	−2.1399
	Spectral-CNN	−0.3227	−0.4452	−0.1331	+0.9735	−1.4741
	EEGConformer	+0.0003	−0.5715	−0.3948	−0.5300	−1.6564
	LaBraM	+0.1224	−0.2024	−0.6981	−0.3168	−1.1738
	CBraMod	+0.1373	−1.4133	−0.3996	+1.2719	+0.0893
PD-RS	TCN	+1.2138	+0.6000	+1.5092	+1.0378	+0.9147
	Spectral-CNN	−4.7945	−3.6099	+4.4668	+4.3609	+2.9105
	MedFormer	+2.7336	−2.1599	−2.2415	+1.8551	−0.7209
	EEGConformer	−1.4966	+5.1956	+1.1215	+2.1065	−3.1646
	LaBraM	+0.3244	−3.8662	−3.5103	−0.3683	+8.8802
	CBraMod	−1.5414	−2.6011	−2.2704	+2.0588	+0.6134

Table 6: **Per-fold means of the content contrast.** One setting per row; each entry is the mean over the 32 seeds of the fold's patient-level macro-F1 difference (percentage points) between training with generated data and with a matched amount of extra real windows. Fold-level means vary in sign inside several settings, and the fold structure differs by cohort, so rows are read independently and are not ranked against one another; the properties that travel with these fold means are stated in the text of this section and in Section 4.5.

of a point, and the standard deviations describe seed-to-seed spread rather than the precision of the mean.

No difference is taken in this appendix; Tables 5 to 8 carry the contrasts and their intervals. This notation holds for all four tables, whose captions name only what is particular to their cohort.

Cohort	Architecture	Utility (pt)	Exposure (pt)	Generated run / real run	Matched real run / real run
ADFTD	TCN	$+2.9432$ $[+1.2975, +4.5889]$	-0.3302 $[-2.0950, +1.4346]$	0.929	1.171
	Spectral-CNN	-0.1665 $[-1.0564, +0.7235]$	-0.1976 $[-1.3266, +0.9314]$	1.283	1.248
	MedFormer	$+23.8410$ $[+20.7423, +26.9396]$	$+0.6633$ $[-1.3168, +2.6434]$	1.293	1.247
	EEGConformer	$+4.4090$ $[+2.6210, +6.1971]$	$+1.1814$ $[-0.9071, +3.2699]$	1.240	1.284
	LaBraM	-1.6833 $[-3.0170, -0.3497]$	-0.4490 $[-1.7454, +0.8474]$	1.340	1.313
	CBraMod	$+0.0874$ $[-1.5271, +1.7019]$	-0.2145 $[-1.1162, +0.6871]$	1.670	1.285
ADHD-WMRI	TCN	$+1.4596$ $[-0.3738, +3.2930]$	$+0.0388$ $[-1.9246, +2.0021]$	1.268	1.156
	Spectral-CNN	$+0.6059$ $[-1.5043, +2.7161]$	$+0.8710$ $[-0.9588, +2.7009]$	1.114	1.074
	MedFormer	$+1.4135$ $[-0.6040, +3.4310]$	-0.0654 $[-2.2847, +2.1540]$	1.333	1.179
	EEGConformer	-2.2227 $[-4.3207, -0.1247]$	$+1.2700$ $[-1.7075, +4.2475]$	1.232	1.165
	LaBraM	-1.3859 $[-2.6960, -0.0757]$	-0.2171 $[-1.3679, +0.9337]$	1.222	1.171
	CBraMod	$+1.8371$ $[+1.0467, +2.6275]$	-0.1179 $[-0.7937, +0.5579]$	1.468	1.250
CAUEEG	TCN	$+1.49$ $[+1.06, +1.92]$	$+0.37$ $[-0.02, +0.76]$	1.187	0.944
	Spectral-CNN	$+0.27$ $[-0.19, +0.72]$	$+0.55$ $[-0.08, +1.18]$	1.179	1.205
	EEGConformer	-0.51 $[-1.54, +0.52]$	$+0.12$ $[-0.78, +1.03]$	1.266	1.183
	LaBraM	-0.50 $[-0.91, -0.09]$	-0.04 $[-0.42, +0.33]$	1.231	1.201
	CBraMod	$+0.35$ $[+0.07, +0.62]$	$+0.41$ $[+0.14, +0.67]$	1.379	1.155
PD-RS	TCN	$+2.22$ $[+1.18, +3.26]$	$+1.17$ $[+0.08, +2.25]$	1.565	1.515
	Spectral-CNN	$+1.0749$ $[-0.0155, +2.1654]$	$+0.4082$ $[-0.7114, +1.5278]$	1.718	1.347
	MedFormer	$+0.74$ $[-0.85, +2.34]$	$+0.85$ $[-0.49, +2.19]$	1.464	1.392
	EEGConformer	$+1.0653$ $[-0.6172, +2.7479]$	$+0.3128$ $[-1.4655, +2.0912]$	1.464	1.532
	LaBraM	$+1.2595$ $[-0.6741, +3.1931]$	$+0.9675$ $[-1.0566, +2.9917]$	1.689	1.627
	CBraMod	-0.0844 $[-0.7168, +0.5480]$	$+0.6637$ $[-0.1537, +1.4812]$	1.562	1.458

Table 7: **Utility and exposure, with the training lengths behind them.** One setting per row, with 95% intervals over 32 seeds, in percentage points of patient-level macro-F1. Utility is the difference between training with generated data and on real data alone; exposure is the difference between training on a matched amount of extra real windows and on real data alone. Utility minus exposure gives the content contrast of Table 5, exactly in every seed (equivalently, utility = content + exposure); entries may not sum exactly at the printed precision. Exposure is reported for transparency only. The last two columns are the ratios of per-run training-update totals (generated against real-only, and matched extra real against real-only); the runs are not length-matched in any setting, so differences in training length are never taken as content conclusions. Rows are read independently and are not ranked against one another.

Setting	Primary result	The secondary metrics, all kept
ADFTD LaBraM	-1.2344 $[-2.5080, +0.0393]$, includes zero	patient-pooled -1.3257 $[-2.4384, -0.2130]$; window-level fold-average -0.9729 $[-1.4435, -0.5023]$; window-level pooled -1.0385 $[-1.4525, -0.6245]$; patient AUROC -1.1771 $[-1.7773, -0.5768]$
ADFTD CBraMod	$+0.3019$ $[-1.0891, +1.6930]$, includes zero	window-level -1.5517 $[-1.8571, -1.2464]$, entirely below zero; patient AUROC $+1.0342$ $[+0.5100, +1.5585]$
ADHD-WMRI CBraMod	$+1.9550$ $[+1.1849, +2.7251]$, above zero	AUROC -1.8661 $[-2.1692, -1.5630]$, the opposite sign
ADHD-WMRI Spectral-CNN	-0.2651 $[-2.0258, +1.4956]$, includes zero	AUROC -3.9554 $[-4.9434, -2.9673]$
CAUEEG TCN	$+1.12$ $[+0.71, +1.53]$, above zero	sample-level fold-average -0.22 $[-0.31, -0.13]$; sample-level pooled -0.21 $[-0.30, -0.13]$; patient-pooled $+1.17$; patient AUROC $+0.86$ $[+0.69, +1.03]$

Table 8: **Where a secondary metric disagrees with the primary.** One setting per row; every value is in percentage points of patient-level macro-F1 except the AUROC entries, which are percentage points of AUROC. The primary result is the content contrast of Table 5; the remaining entries are secondary metrics, reported for transparency only. A secondary metric disagrees with the primary when its interval excludes zero where the primary's includes zero, or when it carries the opposite sign. Both are kept in every row, and neither is chosen over the other. Rows are read independently and are not ranked against one another.

Architecture	Training condition	macro-F1	Accuracy	Bal. acc.	AUROC	F_1^+
TCN	Real only (RO)	0.8063±0.0297	0.8163±0.0228	0.8157±0.0258	0.8989±0.0166	0.8143±0.0200
	Real + generated (CF)	0.8357±0.0296	0.8409±0.0252	0.8388±0.0263	0.9218±0.0144	0.8453±0.0308
	Real + matched real (CT)	0.8030±0.0338	0.8111±0.0301	0.8112±0.0323	0.8942±0.0124	0.8083±0.0287
Spectral-CNN	Real only (RO)	0.7703±0.0197	0.7822±0.0184	0.7742±0.0188	0.8744±0.0077	0.8043±0.0178
	Real + generated (CF)	0.7686±0.0171	0.7812±0.0165	0.7733±0.0167	0.8759±0.0078	0.8054±0.0171
	Real + matched real (CT)	0.7683±0.0245	0.7817±0.0233	0.7723±0.0240	0.8738±0.0093	0.8083±0.0225
MedFormer	Real only (RO)	0.3804±0.0366	0.5548±0.0265	0.5106±0.0230	0.5341±0.0847	0.6681±0.0757
	Real + generated (CF)	0.6188±0.0863	0.6880±0.0507	0.6688±0.0582	0.7940±0.0428	0.7336±0.0714
	Real + matched real (CT)	0.3870±0.0495	0.5577±0.0327	0.5146±0.0309	0.5774±0.0840	0.6772±0.0646
EEGConformer	Real only (RO)	0.6887±0.0379	0.7269±0.0300	0.7149±0.0295	0.8426±0.0183	0.7480±0.0473
	Real + generated (CF)	0.7328±0.0290	0.7567±0.0233	0.7454±0.0252	0.8911±0.0226	0.7883±0.0287
	Real + matched real (CT)	0.7005±0.0363	0.7356±0.0319	0.7220±0.0313	0.8384±0.0278	0.7677±0.0391
LaBraM	Real only (RO)	0.8446±0.0241	0.8514±0.0216	0.8484±0.0226	0.9280±0.0125	0.8513±0.0241
	Real + generated (CF)	0.8278±0.0261	0.8346±0.0221	0.8346±0.0227	0.9156±0.0106	0.8273±0.0251
	Real + matched real (CT)	0.8401±0.0250	0.8486±0.0218	0.8447±0.0231	0.9274±0.0123	0.8493±0.0232
CBraMod	Real only (RO)	0.7690±0.0223	0.7798±0.0193	0.7727±0.0204	0.8420±0.0068	0.7918±0.0190
	Real + generated (CF)	0.7699±0.0336	0.7817±0.0323	0.7723±0.0331	0.8514±0.0099	0.8069±0.0320
	Real + matched real (CT)	0.7668±0.0133	0.7788±0.0116	0.7699±0.0129	0.8411±0.0082	0.7967±0.0126

Table 9: **Raw metric values, ADFTD.** The cohort's six settings, one sub-row per training condition.

Architecture	Training condition	macro-F1	Accuracy	Bal. acc.	AUROC	F_1^+
TCN	Real only (RO)	0.5586±0.0457	0.6140±0.0366	0.5811±0.0404	0.7132±0.0186	0.7058±0.0278
	Real + generated (CF)	0.5732±0.0394	0.6285±0.0289	0.5962±0.0341	0.7460±0.0196	0.7150±0.0211
	Real + matched real (CT)	0.5590±0.0483	0.6175±0.0349	0.5853±0.0416	0.7140±0.0182	0.7084±0.0233
Spectral-CNN	Real only (RO)	0.5838±0.0420	0.6143±0.0361	0.5913±0.0371	0.6010±0.0309	0.6878±0.0291
	Real + generated (CF)	0.5898±0.0428	0.6171±0.0337	0.5968±0.0367	0.5710±0.0173	0.6851±0.0296
	Real + matched real (CT)	0.5925±0.0378	0.6152±0.0316	0.5975±0.0336	0.6105±0.0302	0.6803±0.0250
MedFormer	Real only (RO)	0.6133±0.0451	0.6508±0.0354	0.6287±0.0402	0.7252±0.0265	0.7216±0.0257
	Real + generated (CF)	0.6275±0.0329	0.6637±0.0233	0.6419±0.0274	0.7357±0.0234	0.7320±0.0208
	Real + matched real (CT)	0.6127±0.0423	0.6479±0.0346	0.6291±0.0390	0.7178±0.0275	0.7163±0.0274
EEGConformer	Real only (RO)	0.5558±0.0502	0.6156±0.0352	0.5858±0.0404	0.6410±0.0405	0.6964±0.0272
	Real + generated (CF)	0.5335±0.0363	0.6075±0.0287	0.5735±0.0303	0.6286±0.0337	0.6929±0.0279
	Real + matched real (CT)	0.5685±0.0627	0.6250±0.0450	0.5999±0.0525	0.6288±0.0324	0.6958±0.0344
LaBraM	Real only (RO)	0.6896±0.0220	0.7205±0.0176	0.6911±0.0199	0.7485±0.0179	0.7817±0.0120
	Real + generated (CF)	0.6757±0.0343	0.7062±0.0265	0.6789±0.0292	0.7399±0.0162	0.7629±0.0230
	Real + matched real (CT)	0.6874±0.0232	0.7174±0.0187	0.6877±0.0211	0.7484±0.0220	0.7786±0.0135
CBraMod	Real only (RO)	0.6036±0.0137	0.6714±0.0097	0.6320±0.0099	0.7492±0.0055	0.7582±0.0084
	Real + generated (CF)	0.6219±0.0203	0.6838±0.0162	0.6457±0.0171	0.7307±0.0047	0.7651±0.0103
	Real + matched real (CT)	0.6024±0.0195	0.6661±0.0162	0.6285±0.0165	0.7494±0.0073	0.7518±0.0132

Table 10: **Raw metric values, ADHD-WMRI.** The cohort's six settings; the 11 to 12 subject test folds that qualify its results qualify these values as well.

Architecture	Training condition	macro-F1	Accuracy	Bal. acc.	AUROC	F_1^+
TCN	Real only (RO)	0.7698±0.0100	0.8113±0.0062	0.7544±0.0107	0.8817±0.0044	0.6723±0.0168
	Real + generated (CF)	0.7847±0.0088	0.8217±0.0062	0.7707±0.0092	0.8883±0.0031	0.6957±0.0140
	Real + matched real (CT)	0.7735±0.0079	0.8124±0.0057	0.7595±0.0081	0.8797±0.0036	0.6799±0.0123
Spectral-CNN	Real only (RO)	0.7493±0.0113	0.7957±0.0071	0.7372±0.0122	0.8624±0.0066	0.6425±0.0200
	Real + generated (CF)	0.7520±0.0107	0.7956±0.0071	0.7410±0.0112	0.8609±0.0050	0.6489±0.0181
	Real + matched real (CT)	0.7548±0.0139	0.7979±0.0088	0.7429±0.0144	0.8662±0.0066	0.6526±0.0231
EEGConformer	Real only (RO)	0.7493±0.0208	0.7891±0.0154	0.7460±0.0201	0.8471±0.0145	0.6528±0.0346
	Real + generated (CF)	0.7442±0.0208	0.7877±0.0147	0.7396±0.0212	0.8490±0.0124	0.6416±0.0361
	Real + matched real (CT)	0.7505±0.0205	0.7926±0.0141	0.7451±0.0190	0.8470±0.0123	0.6504±0.0335
LaBraM	Real only (RO)	0.8040±0.0075	0.8411±0.0058	0.7854±0.0074	0.9115±0.0031	0.7191±0.0114
	Real + generated (CF)	0.7991±0.0088	0.8378±0.0061	0.7805±0.0092	0.9100±0.0033	0.7112±0.0142
	Real + matched real (CT)	0.8036±0.0066	0.8402±0.0049	0.7862±0.0069	0.9105±0.0032	0.7192±0.0105
CBraMod	Real only (RO)	0.7634±0.0061	0.8100±0.0042	0.7456±0.0061	0.8767±0.0026	0.6585±0.0100
	Real + generated (CF)	0.7668±0.0049	0.8123±0.0036	0.7494±0.0049	0.8750±0.0022	0.6641±0.0079
	Real + matched real (CT)	0.7675±0.0055	0.8116±0.0039	0.7506±0.0057	0.8758±0.0023	0.6664±0.0089

Table 11: **Raw metric values, CAUEEG.** The cohort's four settings together with the TCN replication run.

Architecture	Training condition	macro-F1	Accuracy	Bal. acc.	AUROC	F_1^+
TCN	Real only (RO)	0.6446±0.0261	0.6869±0.0237	0.6487±0.0260	0.7198±0.0179	0.7646±0.0206
	Real + generated (CF)	0.6668±0.0215	0.7054±0.0202	0.6711±0.0216	0.7235±0.0122	0.7781±0.0174
	Real + matched real (CT)	0.6563±0.0170	0.6948±0.0170	0.6624±0.0153	0.7143±0.0155	0.7684±0.0155
Spectral-CNN	Real only (RO)	0.7230±0.0248	0.7581±0.0199	0.7269±0.0258	0.7938±0.0166	0.8164±0.0193
	Real + generated (CF)	0.7338±0.0192	0.7679±0.0188	0.7384±0.0185	0.7909±0.0095	0.8239±0.0178
	Real + matched real (CT)	0.7271±0.0220	0.7591±0.0218	0.7346±0.0204	0.8020±0.0109	0.8155±0.0207
MedFormer	Real only (RO)	0.6034±0.0262	0.6610±0.0185	0.6084±0.0250	0.6828±0.0159	0.7488±0.0146
	Real + generated (CF)	0.6108±0.0324	0.6546±0.0249	0.6194±0.0320	0.6778±0.0201	0.7352±0.0213
	Real + matched real (CT)	0.6119±0.0301	0.6504±0.0249	0.6220±0.0315	0.6801±0.0191	0.7279±0.0229
EEGConformer	Real only (RO)	0.6719±0.0344	0.7102±0.0306	0.6889±0.0315	0.7633±0.0124	0.7725±0.0303
	Real + generated (CF)	0.6825±0.0287	0.7186±0.0253	0.6955±0.0258	0.7622±0.0165	0.7824±0.0232
	Real + matched real (CT)	0.6750±0.0335	0.7127±0.0284	0.6888±0.0346	0.7570±0.0172	0.7764±0.0278
LaBraM	Real only (RO)	0.6267±0.0399	0.6910±0.0278	0.6329±0.0339	0.7432±0.0172	0.7746±0.0238
	Real + generated (CF)	0.6393±0.0292	0.6981±0.0209	0.6426±0.0300	0.7269±0.0179	0.7803±0.0175
	Real + matched real (CT)	0.6364±0.0336	0.6995±0.0240	0.6401±0.0318	0.7365±0.0189	0.7829±0.0200
CBraMod	Real only (RO)	0.6699±0.0111	0.7221±0.0112	0.6647±0.0104	0.7423±0.0053	0.8002±0.0099
	Real + generated (CF)	0.6691±0.0126	0.7196±0.0116	0.6648±0.0118	0.7435±0.0062	0.7976±0.0099
	Real + matched real (CT)	0.6766±0.0162	0.7225±0.0154	0.6733±0.0158	0.7447±0.0055	0.7972±0.0131

Table 12: **Raw metric values, PD-RS.** The cohort's five settings together with the TCN replication run.