MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Pushing Past Saturation: An EEG Spike-Detection System for NeuroMM-2026 Track-1

Chiang, Ming-Chun; Peng, Kuan-Chuan; Yu, Bo-Yun; Hsieh, Jun-Wei

TR2026-172 November 11, 2026

Abstract

We report the rank-1 system for NeuroMM-2026 Track-1, binary EEG spike detection scored by AUPRC. The system combines domain-aware preprocessing, five time–frequency representations, ImageNet-pretrained vision backbones adapted from prior competition solutions, and a regularized Nelder–Mead ensemble, reaching an inductive score of 0.9711. A conventional pseudo-labelling warm-up improves this to 0.9778, after which our candidate-specialist self-training procedure adds high-confidence released candidates at full weight and reaches 0.9846. Because this final regime is transductive and in-sample by design, we evaluate it with a held-out-fold replay, which suggests partial transfer while not excluding candidate-specific memorization. The main lesson is that, on this saturated EEG benchmark, broad model diversity is insufficient; the decisive lever is many confident pseudo-labelled candidates, full commitment, and candidate-excluded real-label validation. Code: https://github.com/GG3BBE0/PushingPastSaturation-NeuroMM-2026-Track1.

ACM International Conference on Multimedia NeuroMM Challenge 2026

© 2026 ACM. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org or Publications Dept., ACM, Inc., fax +1 (212) 869-0481.

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Pushing Past Saturation: An EEG Spike-Detection System for NeuroMM-2026 Track-1

Ming-Chun Chiang
College of Artificial Intelligence, National Yang Ming Chiao Tung University
Tainan, Taiwan
nycucleglas.ai14@nycu.edu.tw

Kuan-Chuan Peng
Mitsubishi Electric Research Laboratories
Cambridge, MA, USA
kpeng@merl.com

Bo-Yun Yu
College of Artificial Intelligence, National Yang Ming Chiao Tung University
Tainan, Taiwan
boyunyu.ai13@nycu.edu.tw

Jun-Wei Hsieh
College of Artificial Intelligence, National Yang Ming Chiao Tung University
Tainan, Taiwan
jwhsieh@nycu.edu.tw

Abstract

We report the rank-1 system for NeuroMM-2026 Track-1, binary EEG spike detection scored by AUPRC. The system combines domain-aware preprocessing, five time–frequency representations, ImageNet-pretrained vision backbones adapted from prior competition solutions, and a regularized Nelder–Mead ensemble, reaching an inductive score of 0.9711. A conventional pseudo-labelling warm-up improves this to 0.9778, after which our candidate-specialist self-training procedure adds high-confidence released candidates at full weight and reaches **0.9846**. Because this final regime is transductive and in-sample by design, we evaluate it with a held-out-fold replay, which suggests partial transfer while not excluding candidate-specific memorization. The main lesson is that, on this saturated EEG benchmark, broad model diversity is insufficient; the decisive lever is many confident pseudo-labelled candidates, full commitment, and candidate-excluded real-label validation. Code: https://github.com/GG3BBE0/PushingPastSaturation-NeuroMM-2026-Track1.

CCS Concepts

• **Computing methodologies** → **Neural networks**; **Supervised learning by classification**; • **Applied computing** → *Health informatics*.

Keywords

EEG, interictal epileptiform discharge, spike detection, transductive self-training, model ensembling, time–frequency representation, reproducibility

ACM Reference Format:
Ming-Chun Chiang, Kuan-Chuan Peng, Bo-Yun Yu, and Jun-Wei Hsieh. 2026. Pushing Past Saturation: An EEG Spike-Detection System for NeuroMM-2026 Track-1. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3767308.3837695

1 Introduction

Starting from a strong inductive baseline with an AUPRC of 0.9711, a conventional pseudo-labelling warm-up improves performance to 0.9778 (+0.0067), after which the gain saturates. We then introduce a candidate-specialist refinement step, which provides the decisive additional improvement of +0.0068, reaching the final AUPRC of 0.9846 and yielding a total transductive gain of +0.0135. This setting follows the challenge-permitted protocol and is intentionally in-sample with respect to the released candidate set by design (§3.5). Held-out-fold simulations further suggest partial transfer to unseen labelled folds, although they do not fully exclude the possibility of candidate-specific adaptation (§3.6).

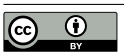
We emphasize that the individual model architectures are adapted from prior work; therefore, this paper should be understood primarily as a rigorous winning-system report rather than a purely architecture-centric contribution. Nevertheless, the proposed system contributes a new integrated framework that combines strong EEG representation models, pseudo-labelling, and candidate-specialist refinement into an effective and reproducible pipeline. Its value lies in system-level integration, careful empirical validation, reproducibility, and practical lessons for automated EEG interpretation under realistic challenge settings.

Contributions:

(1) **A rank-1 reproducible EEG-spike detection system.** We develop an end-to-end system that combines domain-aware EEG preprocessing, five complementary time–frequency representations, adapted ImageNet-pretrained backbones, and a regularized Nelder–Mead ensemble optimization strategy. The resulting system achieves rank 1 on the official leaderboard.

(2) **A candidate-specialist self-training analysis under inductive saturation.** We conduct a systematic study of self-training in a saturated regime and show that, after a strong inductive baseline has plateaued, using many high-confidence candidates at full weight is more effective than sparse, down-weighted pseudo-labelling. Held-out-fold replay experiments and ablation studies further characterize the source of the gain, its robustness, and its limitations.

We are explicit about attribution and scope. Our stacked-spectrogram CNN follows the channel-stacking scheme of a public HMS solution [30], whereas our multi-view model is a simplified re-implementation of another [25] (§3.3); both use ImageNet-pretrained `timm` [37] backbones. Pseudo-labelling, self-training, and transductive learning are established techniques [8, 16, 38]. The contribution lies in integrating these components into a reproducible rank-1 system, identifying when candidate-specialist self-training remains useful after inductive saturation, and



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3837695

providing an empirical study of its benefits and limitations in EEG-spike detection.

2 Related Work

2.1 EEG spike detection and representations

Automated IED detection has been approached with raw-signal and time–frequency CNNs [13], semi-supervised anomaly models [1], and multimodal detectors that pair EEG with video [20]; in parallel, large EEG foundation models are pretrained on unlabelled recordings (LaBraM [12], BIOT [39], REVE [5]) and recent work couples EEG features to language models for seizure detection [19]. We take the time–frequency-image route and, rather than committing to a single basis, combine complementary ones—Morlet and band-passed CWT, the Paul wavelet [32], superlets [24] and STFT—each fed to ImageNet-pretrained backbones. The resulting ensemble is purely inductive and, as §4.1 shows, it saturates; §2.2 positions the transductive stage that breaks it.

2.2 Semi-supervised and transductive self-training

Pseudo-labelling [16], Noisy Student [38] and learning by transduction [8] are the classical anchors, and modern deep semi-supervised learning is broad [40]. Our candidate-specialist stage occupies a narrow position in that space. The unlabelled set is not auxiliary data: it is the released candidate pool that is itself scored. A converged ensemble teacher assigns hard labels to a large tail-selected subset, and those candidates enter training at full weight. Each teacher iteration is accepted only if cross-validation computed purely on real labels improves, which is well defined because every pseudo-labelled candidate is tagged fold=−1 and never enters a validation fold (§3.5). The challenge test labels are hidden and the leaderboard is closed, so re-running the ladder under each alternative design is not possible. We therefore position this design against four recent families analytically rather than empirically.

(1) Clustering-first and incremental self-training. Incremental self-training [9] clusters the unlabelled pool first and measures certainty by the Euclidean distance from a sample to its cluster centre, admitting confident samples early and decision-boundary samples only once the model has stabilized. Clustering-based SSL [7] instead ties cluster prototypes to class prototypes and learns assignments and labels jointly under a single objective. We never cluster in feature space: the inclusion criterion is the teacher posterior, gated externally by real-label cross-validation. This avoids assuming that EEG windows form clean clusters, an assumption strained by the inter-subject and inter-montage variability of scalp EEG. A posterior-based criterion also stays comparable across heterogeneous backbones whose feature spaces are not directly comparable. What we give up is the curriculum those methods provide implicitly: without easy-first ordering and a second look at boundary samples, a candidate the teacher misjudges is not re-examined within a round.

(2) Disagreement-weighted and mutual training. DMT [6] co-trains two differently initialized models and down-weights the pseudo-label loss wherever they disagree. The related co-teaching [10] exchanges small-loss samples between peers, though as a noisy-label-robustness device rather than a semi-supervised disagreement scheme. We do not modulate the loss at all. High-confidence candidates are committed at weight 1.0, and the risk is carried by an external gate, candidate-excluded real-label CV, instead of an internal re-weighting term. §4.2(B) shows that at saturation down-weighting caps the achievable gain whereas full commitment raises every retrained specialist, and one external gate is cheaper than co-training and coordinating several students. What we forgo is a cross-model signal for errors the teacher makes confidently and consistently: the gate can reject an entire round, but not an individual mislabelled candidate.

(3) Layer-specific gating and pseudo-label routing. LayerMatch [18] separates the effect of pseudo-labels across network depth. Its Grad-ReLU component suppresses the pseudo-label gradient at the linear classifier while the feature extractor still receives it, and its Avg-Clustering component stabilizes the cluster structure that extractor learns, so noisy pseudo-labels may shape representations without moving the decision boundary. Our gating acts at the sample level rather than the network level: a candidate is either admitted at full weight or discarded, and once admitted the ordinary loss reaches all trainable parameters. Sample-level gating is architecture-agnostic, so one recipe transfers unchanged across our pool of CNN and transformer backbones. Those backbones share no common layer taxonomy, so a depth-wise rule could not be defined consistently across them. The cost is coarser control: label noise that survives the threshold reaches the classifier as well as the features.

(4) EEG-specific semi-supervised and test-time adaptation. Semi-supervised learning for EEG has been reviewed primarily for emotion recognition [29], with representative methods such as multimodal semi-supervised affect recognition [3], while self-supervised EEG pretraining is surveyed separately [36]. For clinical IED analysis, prior semi-supervised methods use autoencoder-based anomaly modelling [1] or temporal-autoencoder clustering of intracranial EEG from few gold labels [26]. A related deployment-time direction is test-time adaptation: T-TIME adapts an ensemble to streaming unlabelled target-subject EEG by information maximization [17], TENT provides the general entropy-minimization framework [34], and NeuroTTT extends test-time training to EEG foundation models in a recent preprint [35]. Transductive protocols remain rare in EEG, where the unlabelled pool is normally treated as an annotation-cost problem rather than as scored data. Those lines use unlabelled EEG to cut annotation cost or to adapt to unseen subjects. We instead retain complete real-label supervision and specialize offline to a fixed released pool that is itself the scored set. Direct access to the released candidates lets the specialists adapt to the target input distribution, which reduces the source–target mismatch that often limits transfer in EEG. In exchange we inherit none of the EEG-specific consistency, augmentation or subject-adaptation priors those lines develop, and we take on a stronger candidate-specific adaptation risk.

Scope of the transductive setting. The final gain is by construction tied to a fixed released candidate set: a transductive learner improves predictions on the very samples it trains on, so a leaderboard number alone cannot separate genuine generalization from candidate-specific adaptation. We therefore rely on the held-out-fold replay of §4.3, which we read conservatively: it repeats the procedure on labelled folds the teacher never saw and is positive

on 4/5 folds, which is suggestive of partial transfer but does not exclude memorization of the particular candidates. Families (1)–(3) are normally evaluated under inductive protocols, although they too could consume the released pool as unlabelled training data. The distinction is not whether the pool may be used, but whether training is explicitly specialized to improve predictions on that same fixed evaluation pool. We regard this as a property of the evaluation protocol rather than a defect of the procedure.

3 Method

The system is an inductive pipeline (§3.2–§3.4) that saturates, followed by a transductive self-training step (§3.5) that breaks saturation, plus a generalization check (§3.6). Figure 1 gives the overview.

3.1 Problem and notation

We have N=25,426 labelled windows from the NeuroMM-2026 Track-1 corpus [21, 31] under a 5-fold *patient-disjoint* split, with overall positive rate 9.9% (per-fold 5.8%–13.0%), and U=20,000 unlabelled candidates that constitute the scored set. We write AUPRC_{cv} for pooled 5-fold out-of-fold AUPRC on real labels, and denote by fold=−1 a sample that is always in training and never in validation.

3.2 Preprocessing and representations

Given the class imbalance, a recurring obstacle in clinical EEG classification [4], we use focal loss [22] as a standard choice for hard-example emphasis. Each 4-second window (2000 samples at 500 Hz) is reduced from 29 to 26 channels by scaling EEG and physiological channels to comparable ranges and forming three bipolar derivations (ECG and two EMGs), which is a standard montage choice. A *band-passed* branch additionally applies a 4th-order Butterworth band-pass (0.5–70 Hz, zero-phase), a 50 Hz notch, an EEG-only common-average reference, and a per-channel robust normalization $(x-\text{median})/(\text{IQR}+\epsilon)$. On the *unfiltered* montage signal we compute raw EEG, Morlet CWT, Paul-wavelet CWT [32], the superlet [24] (geometric mean of |CWT| over Morlet cycle counts 3–7, a batched-FFT GPU kernel), and an STFT; the band-passed branch yields band-passed EEG and BP-CWT (the CWT of that band-passed signal). This gives two time-domain inputs (26×2000) and five complementary time–frequency representations (wavelet maps 26×64×256, STFT 26×65×59). Adapting the channel-stacking scheme of a public HMS solution [30], a multi-channel spectrogram $x \in \mathbb{R}^{C\times F\times T}$ (channels × frequency × time) is stacked along frequency into a single-channel image, bilinearly resized to $S \times S$ ($S \in \{256, 384\}$), and fed to an ImageNet-pretrained backbone; we write `rep@S` throughout for a representation `rep` at input resolution S×S.

3.3 Diverse model pool

We instantiate the adapted stacked-spectrogram CNN across five representations, two resolutions, and several `timm` backbones (e.g., MaxViT [33] and ConvNeXt [23]); our contribution here is the *breadth* of the instantiation. Inspired by a multi-view EEG model [25], our re-implementation uses a learnable temporal convolution: a one-second kernel filters along the time axis, replacing hand-crafted band-pass filtering with a learned filter bank (LFB). Its output is treated as a time–frequency map and reordered into two complementary views, channel-major and frequency-major, each fed to its own backbone. Relative to the source, we use only a single LFB rather than its multi-scale stack, and drop its spectrogram and recurrent branches—a simplification suited to the smaller NeuroMM data. Following common deep-supervision practice, these models carry auxiliary heads; our specific choice is to apply the focal loss only to the main head while keeping binary cross-entropy (BCE) on the auxiliary heads, so the focal gain remains attributable to the main head. Finally, a few legacy time-domain EEG CNNs (an EEGNet-style compact CNN [15] and a MobileNet-based model [11]) from early experiments are kept in the pool for diversity; they carry negligible final weight (Table 1).

Representation analysis. On the only same-backbone comparison level (384 px), Paul is the single strongest representation (0.858), with superlet second (0.849) and CWT third (0.841); their 95% bootstrap intervals overlap heavily, so *no single representation dominates*. The superlet member earns the largest ensemble weight only after heavy augmentation and at 384 px (§4.2), not because the representation is intrinsically superior.

3.4 Regularized Nelder–Mead ensemble

We combine member out-of-fold logits by maximizing AUPRC over non-negative simplex weights with Nelder–Mead [27]. Because AUPRC is non-smooth in the weights, gradient and SLSQP [14] solvers collapse to a near-uniform (simple-average) solution. Therefore, a derivative-free search is required. Three regularizers stabilize the high-dimensional weight estimate: *(i) Two-dimensional trimming:* a member is dropped only if it is *both* individually weak (single-model $\text{AUPRC}_{\text{cv}} < 0.68$) *and* assigned near-zero weight (< 0.005)—either criterion alone is unreliable, since a weak member can still earn weight through useful decorrelation and a strong one can be redundant with the rest. *(ii) Bagging* [2]: we average the weights over bootstrap resamples of the validation set, damping the variance of any single non-convex optimum (reruns then differ only marginally; the spread is quantified in §4). *(iii) Capping* per-member weight at 0.20 (a conservative choice; sweeping it barely moves the in-sample AUPRC_{cv}). The resulting ensemble of 43 members (reruns vary 43–45) reaches an in-sample AUPRC_{cv} of 0.9269 and a leaderboard 0.9711. These weights are fit on the same AUPRC_{cv}, so 0.9269 is optimistic—yet it still sits *below* the leaderboard; the more conservative nested-CV estimate, with weights fit on disjoint folds, is reported in §4.2. Adding further same-family architectures inflates in-sample AUPRC_{cv} without improving the leaderboard—a symptom of saturation—which motivates a different lever.

3.5 Candidate-specialist transductive self-training

Rather than adding correlated architectures, we exploit the 20k unlabelled candidates as in-distribution training data. This is permitted by the challenge and uses *no hidden labels*; the strongest available ensemble teacher assigns a pseudo-label to each *confident* candidate—a high-probability tail above a threshold HI becomes a pseudo-positive, a low-probability tail below LO a pseudo-negative, and the uncertain middle is discarded so that only labels the teacher is sure about enter training.

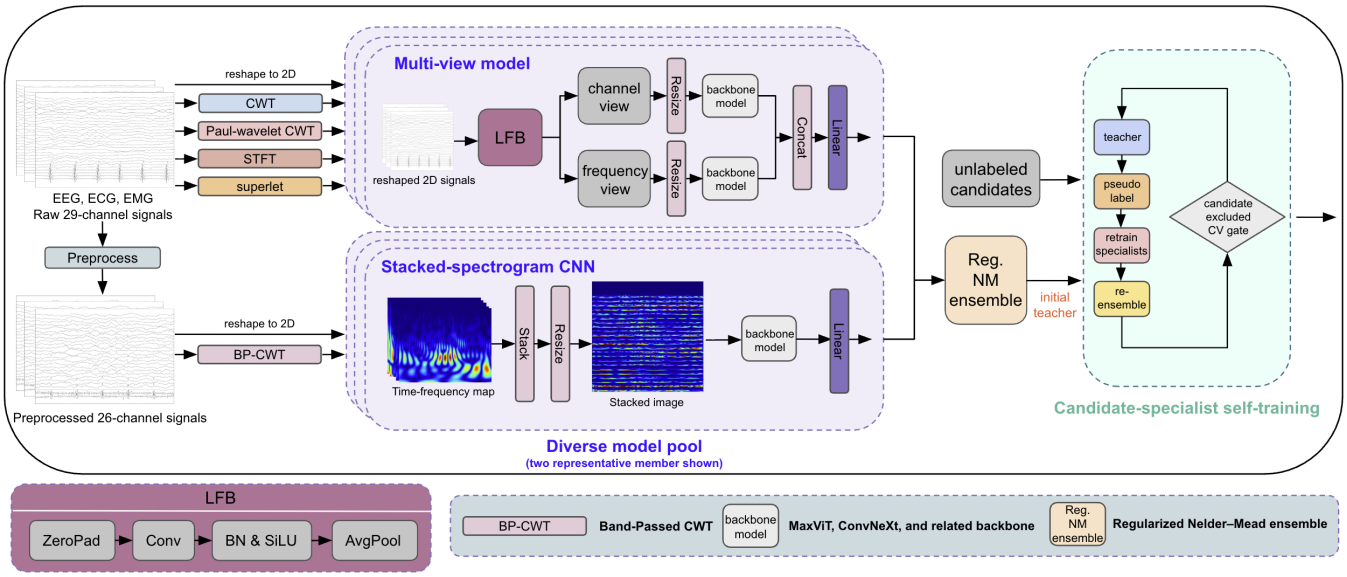


Figure 1: System data-flow. An inductive pipeline—preprocessing, five time–frequency representations, a diverse model pool (two representative members shown), and a regularized Nelder–Mead ensemble—is broken past saturation by candidate-specialist transductive self-training, whose teacher iteration is gated by candidate-excluded cross-validation. Backbones are shown as boxes, not internal architectures: the spectrogram members use off-the-shelf `timm` backbones [30, 37], while the multi-view model is our simplified re-implementation of a public design [25].

The fold=−1 real-label gate. Every pseudo-labelled candidate is tagged fold=−1, meaning it is always in the training partition and never in any of the five validation folds. Consequently the cross-validation that decides whether to accept a self-training round is computed entirely on *real-labelled* held-out folds that contain no pseudo-labels. This makes the gate a conservative real-label check against self-amplification: if pseudo-label error amplification also damages transferable decision structure, it is expected to appear as a *drop* on the real-labelled folds rather than as a spurious gain. Accepting a round only when this real-label CV rises beyond the noise floor therefore lets us train candidates at full weight while providing a round-level check; it cannot, however, detect candidate-specific errors that leave the real-label CV unchanged.

An initial pseudo-labelling phase (a few thousand down-weighted candidates) reaches 0.9778 (+0.0067) and saturates. Our candidate-specialist procedure (Algorithm 1, two teacher iterations) then takes this to 0.9818 and 0.9846 (+0.0068 in total, the second contribution to the +0.0135 transductive gain). Four design choices matter: (i) *quantity*—a permissive threshold yields 11,336 candidates (4,336 positive, 7,000 negative), several times the pseudo phase; (ii) *full commitment*—candidate weight 1.0, reversing the pseudo phase's down-weighting (0.4–0.5); (iii) the fold=−1 real-label gate; and (iv) teacher iteration (the second teacher is the first specialist ensemble). The single-model real-label AUPRC_{cv} of every one of the five retrained specialists rises—+0.036 to +0.080 (Table 3)—yet we note that the final 0.9818 → 0.9846 leaderboard step is an *ensemble*/teacher-iteration effect, not a further rise in those single-model AUPRC_{cv} values: the stronger second teacher selects cleaner pseudo-labels and the recalibrated specialists recombine slightly better under Nelder–Mead. We flag this final rung as the least mechanistically resolved step of the ladder.

Algorithm 1: Candidate-specialist transductive self-training (run for 2 rounds).

Input: labelled $\mathcal{D}$, unlabelled candidates U, thresholds (HI, LO)
$E_0 \leftarrow$ NM-Ensemble(model pool);
for $t = 0, 1$ **do**
  $P_t \leftarrow \{(p, E_t(p)) : p \in U,\ E_t(p) \geq \text{HI} \text{ or } E_t(p) \leq \text{LO}\}$;
  $\mathcal{D}_t \leftarrow \mathcal{D} \cup \{(p, \hat{y}, \text{fold}{=}{-1},\ w{=}1) : (p, \hat{y}) \in P_t\}$;
  retrain specialists on $\mathcal{D}_t$ with focal loss;
  $E_{t+1} \leftarrow$
   NM-Ensemble(model pool with specialists swapped in);
  accept iff real-label AUPRC_{cv} rises beyond the noise floor;

3.6 Held-out-fold generalization check

Because the leaderboard gain is in-sample, we run a cross-validation-style check (not a new method). For each fold k, a teacher trained only on the other four folds pseudo-labels fold k; we inject those pseudo-labels at fold=−1, retrain a specialist, and evaluate on fold k's *true* labels (Figure 2). This faithfully replays the transductive procedure on labelled held-out data, with fold k playing the role of the candidate set. *Limitation:* it is a train-fold proxy and cannot directly rule out memorization of the specific released candidates.

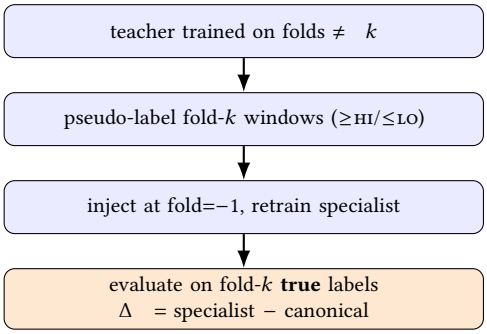


Figure 2: Held-out-fold generalization check: the transductive procedure replayed on labelled held-out data, with fold-k in the role of the candidate set. Repeated for each k; positive Δ on unseen true labels is suggestive of partial transfer of the self-training mechanism.

Table 1: Model-pool composition. The *NM wt.* column (per-family ensemble weight, summing to 1) splits into the self-training *specialist* share and the inductive-pool *other*; the five specialists carry 0.649. multi-view and legacy are time-domain. Weights derived in §3.4.

Family	Members	NM wt.	specialist	other
multi-view	14	0.252	0.159	0.093
superlet	3	0.251	0.200	0.051
BP-CWT	4	0.176	0.143	0.033
Paul	5	0.154	0.091	0.063
CWT	13	0.145	0.056	0.089
legacy	2	0.013	—	0.013
STFT	2	0.009	—	0.009
Total	43	1.000	0.649	0.351

4 Experiments

Setup. 5-fold patient-disjoint CV; AUPRC. Inductive models were trained on RTX 3090 GPUs and the final specialists on an HPC cluster with H100 GPUs; the released system rebuilds on CPU. Single-model scores are pooled 5-fold AUPRC_{cv} with bootstrap 95% intervals; ensemble comparisons use paired multi-seed differences. The final 0.9846 ensemble has 43 members after trimming; the five candidate specialists (superlet@384, multi-view, and CWT/Paul/BP-CWT@256) carry ≈65% of the weight, the superlet specialist alone hitting the 0.20 cap, with the remaining weight spread across the diverse model pool (Table 1 gives the composition by representation).

Ensemble stability and reseeding sensitivity. Re-deriving the bagged Nelder–Mead weights under eight bagging seeds leaves the pooled AUPRC_{cv} near-identical (mean 0.9270, SD 0.0002) and the leaderboard within 0.9842–0.9852 (mean 0.9847); the late ladder steps clear this band, so the gains are not optimizer noise. Significance against the hidden test labels remains out of reach, so the central claim leans on the multi-seed held-out-fold evidence (§4.3).

Table 2: Score ladder on the official test set; stages from 0.9778 are transductive.

Score	Method	Δ	Regime
0.9247	15-arch baseline	—	inductive
0.9446	full pool (23 arch)	+0.0199	inductive
0.9605	+STFT/Paul views (29)	+0.0159	inductive
0.9663	+band-passed EEG (33)	+0.0058	inductive
0.9711	regularized NM ensemble	+0.0048	inductive (sat.)
0.9778	pseudo-labelling phase	+0.0067	transductive
0.9818	candidate specialist (full wt.)	+0.0040	transductive
0.9846	specialist + teacher iter.	+0.0028	transd. (rank 1)

4.1 Main result

Table 2 traces the ladder from a 15-architecture baseline (0.9247) through inductive saturation (0.9711) to 0.9846. The stages from 0.9778 are *transductive*—the specialists train on pseudo-labelled released candidates (~57% of the scored set, challenge-permitted)—so 0.9846 carries an in-sample component, and out-of-sample generalization is assessed separately in §4.3.

4.2 Ablations

(A) Augmentation is representation-dependent. At a fixed backbone, heavy SpecAugment [28] helps superlet@384 (+0.0167), Paul@256 (+0.008) and CWT@256 (+0.0036) but *hurts* BP-CWT@256 (−0.003, the input is already denoised). This, with the 384 px resolution, explains the superlet member's large ensemble weight—not representation superiority.

(B) Full commitment to many candidates, not a tuned threshold. Candidate self-training and the earlier pseudo-labelling are the *same* procedure with opposite hyper-parameters: the pseudo phase treats pseudo-labels as noise (few candidates, down-weighted 0.4–0.5), spanning 2,695 at threshold 0.90 up to 3,438 at threshold 0.80/weight 0.5, and all stall at 0.978; the candidate specialist commits fully (11,336 at threshold 0.60, weight 1.0). The two regimes separate at *two* levels.

Single-model OOF (Table 3). Down-weighting yields *inconsistent* gains—round-1 $\{+0.014, -0.006, -0.003\}$ and round-3 $\{-0.010, +0.042, +0.020\}$ on CWT/Paul/BP-CWT, with some members regressing whereas full weight raises *every* one of the five retrained specialists *uniformly*, by +0.036 to +0.080 (CWT/Paul/BP-CWT/superlet and multi-view).

Ensemble (nested real-label CV). Swapping the full-weight specialists into the ensemble raises its nested out-of-fold AUPRC by **+0.0386** over the inductive baseline (all five seeds positive, +0.032 to +0.043; +0.0131 over the pseudo-labelled system). This is computed on held-out *real* folds that never contain candidates, so it is a genuine out-of-fold gain, not in-sample fitting. Full-weight training is made more defensible by the fold=−1 injection, which keeps the validation folds free of pseudo-labels: confirmation bias that damages transferable decision structure is expected to appear as a CV drop, although candidate-specific errors that leave the real-label CV unchanged may remain undetected. **Lesson:** in this saturated setting, committing fully to many candidates was more effective

Table 3: Per-specialist single-model self-training gain (real-label $\mathrm{AUPRC}_{\mathrm{cv}}$, specialist − base, pooled 5-fold); all five improve, the consistency being the evidence for the quantity×commitment lesson.

Specialist	self-training Δ
CWT@256	+0.0372
Paul@256	+0.0474
BP-CWT@256	+0.0361
superlet@384	+0.0587
multi-view[†]	+0.0800

[†]time-domain; lower base, more headroom.

than conservative down-weighting at any threshold we tried, provided each round was audited with candidate-excluded CV. We read +0.0386 for *direction*, not magnitude: it over-estimates the leaderboard +0.0135 roughly threefold (CV runs on noisier labels with more headroom while the cleaner candidates are nearer ceiling), so the evidence we rely on is the consistently positive sign across all five seeds, and we use this CV as a directional accept/reject gate, not a calibrated predictor of test improvement.

(C) Hard targets outperform soft targets at saturation. A soft-target self-training variant, in the spirit of [38], gains only +0.0020 nested $\mathrm{AUPRC}_{\mathrm{cv}}$ (within the ±0.005–0.01 noise floor), losing the quantity edge of hard pseudo-labels.

4.3 Generalization check

On the held-out-fold check (§3.6), per-fold Δ is positive on 4/5 folds for both CWT (mean +0.0318) and multi-view (mean +0.0547). The CWT mean is outlier-driven (fold 2, +0.125), so we foreground the consistently positive *sign*—suggestive of partial transfer, not decisive. A stronger teacher yields a larger gain (+0.0318 vs. +0.0022 for a weak single-model teacher). Fold 4 is the only non-positive fold in both.

5 Discussion

Why architectural diversity stopped paying. The inductive half of Table 2 improves in steps that shrink as the pool grows (+0.0199, +0.0159, +0.0058, +0.0048), and §3.4 indicates why: the ensemble weights are fitted on the same out-of-fold predictions used to report the score, so in-sample $\mathrm{AUPRC}_{\mathrm{cv}}$ can keep rising while the leaderboard does not. Past a certain pool size the quantity that overfits is no longer any individual model but the weight vector, so a candidate architecture has to be judged on an estimate whose weights were fitted on disjoint folds; on the in-sample figure alone, every addition looks like progress.

What the transductive step changed. Once diversity saturates, the binding constraint is no longer capacity or decorrelation but the amount of labelled data available in the distribution being scored. Candidate self-training relaxes that particular constraint: it converts a strong teacher into many more training examples drawn from the scored pool, which is why a gain appears without any new architecture. This also offers a reading of the asymmetry in §4.2(B). Down-weighting keeps the effective number of training examples close to the labelled set and so cannot relieve the constraint, whereas full commitment can.

When the recipe applies. The procedure needs the scored set to exist before training. That holds in a challenge and in retrospective review of an archived corpus, but not in prospective monitoring where recordings arrive continuously. In that regime the analogue is not candidate specialization but the test-time adaptation line of §2.2, which updates a deployed model on unlabelled data as it arrives. The two are complementary: when a fixed pool is available and a strong teacher exists, committing to that pool was worth more than any further architectural diversity we found.

What the comparison amounts to. Seen against the four families of §2.2, our design trades machinery for a single external check: no feature-space clustering or admission curriculum, no co-trained peers for a disagreement signal, and no depth-wise routing of the pseudo-label gradient—only a sample-level accept/discard decision whose risk is absorbed by a real-label gate. That trade is what makes full commitment affordable and is equally where the method is weakest: the gate can veto a round but never an individual mislabelled candidate, so a confidently and consistently wrong teacher goes unchallenged.

6 Conclusion

On this saturated benchmark, the effective lever is disciplined transductive self-training—many confident candidates at full commitment, gated by candidate-excluded CV. Quantity×commitment paid off here because three conditions held at once—the teacher was already strong, the unlabelled pool was in-distribution and released in full, and a candidate-excluded split existed to audit every round. Where any of them fails—a weaker teacher, a shifted pool, or no clean real-label holdout—the alternatives we forgo (down-weighting, or a cross-model disagreement signal) become the safer choice. Our evidence is correspondingly conservative: the held-out-fold replay is positive on 4/5 folds and we read it as suggestive of partial transfer, not as proof that the gain is free of candidate-specific adaptation. We contribute a rank-1 reproducible system, a transferable self-training lesson with its scope made explicit, and a map of the saturation frontier. Future work: testing the lesson on other datasets and recovering the per-candidate control a single external gate gives up.

References

[1] Ana Maria Amaro de Sousa, Michel J. A. M. van Putten, Stéphanie van den Berg, and Maryam Amir Haeri. 2024. Detection of Interictal epileptiform discharges with semi-supervised deep learning. *Biomedical Signal Processing and Control* 88 (2024), 105610.

[2] Leo Breiman. 1996. Bagging Predictors. *Machine Learning* 24, 2 (1996), 123–140.

[3] Zebang Cheng, Yuxiang Lin, Zhaoru Chen, Xiang Li, Shuyi Mao, Fan Zhang, Daijun Ding, Bowen Zhang, and Xiaojiang Peng. 2023. Semi-supervised multimodal emotion recognition with expression MAE. In *ACM Multimedia*.

[4] Dalila Cherifi, Tinhinene Ahmed-Zaid, Sarah Boukmouri, and Larbi Boubchir. 2025. Evaluating Class Imbalance Techniques in EEG-Based Machine Learning Models for Parkinson's Disease Detection. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 5817–5821.

[5] Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. 2025. REVE: A foundation model for EEG—adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, Vol. 38.

[6] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. 2022. DMT: Dynamic Mutual Training for Semi-Supervised Learning. *Pattern Recognition* 130 (2022), 108777.

[7] Enrico Fini, Pietro Astolfi, Karteek Alahari, Xavier Alameda-Pineda, Julien Mairal, Moin Nabi, and Elisa Ricci. 2023. Semi-Supervised Learning Made Simple with Self-Supervised Clustering. In *CVPR*. 3187–3197.

[8] Alexander Gammerman, Volodya Vovk, and Vladimir Vapnik. 1998. Learning by Transduction. In *UAI*.

[9] Jifeng Guo, Zhulin Liu, Tong Zhang, and C. L. Philip Chen. 2024. Incremental Self-training for Semi-supervised Learning. *arXiv preprint arXiv:2404.12398* (2024).

[10] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels. In *NeurIPS*.

[11] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, and Vijay Vasudevan. 2019. Searching for MobileNetV3. In *ICCV*.

[12] Weibang Jiang, Liming Zhao, and Bao-Liang Lu. 2024. Large Brain Model for Learning Generic Representations with Tractable EEG Data. In *ICLR*.

[13] Jin Jing, Haoqi Sun, Jennifer A Kim, Aline Herlopian, Ioannis Karakis, Marcus Ng, Jonathan J Halford, Douglas Maus, Fonda Chan, Marjan Dolatshahi, et al. 2020. Automated Detection of Interictal Epileptiform Discharges from Scalp Electroencephalograms by Convolutional Neural Networks. *International Journal of Neural Systems* 30, 4 (2020), 2050030.

[14] Dieter Kraft. 1988. *A software package for sequential quadratic programming*. Technical Report DFVLR-FB 88-28. DLR German Aerospace Center.

[15] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. EEGNet: A Compact Convolutional Neural Network for EEG-based Brain–Computer Interfaces. *Journal of Neural Engineering* 15, 5 (2018), 056013.

[16] Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*.

[17] Siyang Li, Ziwei Wang, Hanbin Luo, Lieyun Ding, and Dongrui Wu. 2024. T-TIME: Test-Time Information Maximization Ensemble for Plug-and-Play BCIs. *IEEE Transactions on Biomedical Engineering* 71, 2 (2024), 423–432.

[18] Chaoqi Liang, Guanglei Yang, Lifeng Qiao, Zitong Huang, Hongliang Yan, Yunchao Wei, and Wangmeng Zuo. 2024. LayerMatch: Do Pseudo-labels Benefit All Layers? *arXiv preprint arXiv:2406.14207* (2024).

[19] Zi Liang, Peng Hu, Lian Li, Zebang Cheng, Yisu Dong, Haibo He, Qiang Lu, and Nan Lin. 2025. EEG-VL: Integrating Visual Features with Large Language Models for Automated Seizure Detection. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*.

[20] Nan Lin, Weifang Gao, Lian Li, Junhui Chen, Zi Liang, Gonglin Yuan, Heyang Sun, Qing Liu, Jianhua Chen, Liri Jin, et al. 2024. vEpiNet: A multimodal interictal epileptiform discharge detection method based on video and electroencephalogram data. *Neural Networks* 175 (2024), 106319.

[21] Nan Lin, Mengxuan Zheng, Lian Li, Peng Hu, Weifang Gao, Heyang Sun, Chang Xu, Gonglin Yuan, Zi Liang, Yisu Dong, et al. 2025. An EEG dataset for interictal epileptiform discharge with spatial distribution information. *Scientific Data* 12, 1 (2025), 229.

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. In *ICCV*.

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *CVPR*.

[24] Vasile V Moca, Harald Bârzan, Adriana Nagy-Dăbâcan, and Raul C Mureşan. 2021. Time-frequency super-resolution with superlets. *Nature Communications* 12, 1 (2021), 337.

[25] muku. 2024. HMS Harmful Brain Activity Classification: solution writeup (multi-view EEG models). Kaggle competition solution. https://www.kaggle.com/competitions/hms-harmful-brain-activity-classification/writeups/team-sony-1st-place-solution-team-sony#2744634.

[26] Petr Nejedly, Vaclav Kremen, Kamila Lepkova, Filip Mivalt, Vladimir Sladky, Tereza Pridalova, Filip Plesinger, Pavel Jurak, Martin Pail, Milan Brazdil, Petr Klimes, and Gregory Worrell. 2023. Utilization of temporal autoencoder for semi-supervised intracranial EEG clustering and classification. *Scientific Reports* 13, 744 (2023).

[27] John A Nelder and Roger Mead. 1965. A simplex method for function minimization. *Comput. J.* 7, 4 (1965), 308–313.

[28] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*.

[29] Sen Qiu, Yongtao Chen, Yulin Yang, Pengfei Wang, Zhelong Wang, Hongyu Zhao, Yuntong Kang, and Ruicheng Nie. 2024. A review on semi-supervised learning for EEG-based emotion recognition. *Information Fusion* 104 (2024), 102190.

[30] suguuuuu. 2024. HMS Harmful Brain Activity Classification: solution writeup (spectrogram stacking). Kaggle competition solution. https://www.kaggle.com/competitions/hms-harmful-brain-activity-classification/writeups/team-sony-1st-place-solution-team-sony#2744628.

[31] Qi Tian, Fei Ma, Haibo He, Laizhong Cui, Nan Lin, Zheng Lian, Philippe Fournier-Viger, Zebang Cheng, et al. 2026. NeuroMM 2026: Interictal Epileptiform Discharge Detection and Localization in Multimodal Neuro-Signals. In *NeuroMM Grand Challenge, ACM Multimedia*.

[32] Christopher Torrence and Gilbert P Compo. 1998. A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society* 79, 1 (1998), 61–78.

[33] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. MaxViT: Multi-axis Vision Transformer. In *ECCV*.

[34] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*.

[35] Suli Wang, Yangshen Deng, Zhenghua Bao, Xinyu Zhan, and Yiqun Duan. 2025. NeuroTTT: Bridging Pretraining–Downstream Task Misalignment in EEG Foundation Models via Test-Time Training. *arXiv preprint arXiv:2509.26301* (2025).

[36] Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Zhaohua Yang, Yuchen Liu, and Yiqiang Chen. 2025. Self-supervised Learning for Electroencephalogram: A Systematic Survey. *Comput. Surveys* (2025).

[37] Ross Wightman. 2019. PyTorch Image Models. https://github.com/huggingface/pytorch-image-models.

[38] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with Noisy Student improves ImageNet classification. In *CVPR*.

[39] Chaoqi Yang, M Brandon Westover, and Jimeng Sun. 2023. BIOT: Biosignal Transformer for Cross-data Learning in the Wild. In *NeurIPS*.

[40] Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. 2023. A Survey on Deep Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2023), 8934–8954.