WHEN COARSE ANNOTATIONS CONCEAL NOVEL STATES: TC-REJECT FOR AUDITING TIME-SERIES LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Coarse interval labels can conceal changes in behavior or physiology, contaminating scientific analyses and downstream models. Auditing such labels is difficult because an unseen state may appear typical of the training distribution while conflicting with its assigned label. In controlled simulations, global atypicality and label agreement fail in opposite conditions: density scores invert for coherent lower-variance novelty, while large scale shifts reverse their relative utility. We therefore introduce **Typicality-Corrected Rejection (TC-REJECT)**, which calibrates complementary multichannel representations and combines them using e-value summation or rank averaging without requiring examples of the target novel state. Across wearable activity, naturalistic electrocorticography, and clinical local field potential recordings from deep-brain nuclei in two participants with deep-brain stimulation, TC-REJECT detects withheld or concealed states and supports selective label review. On prespecified previously unused neural recordings, rank-based fusion is more robust than e-value summation and achieves the highest mean performance among independent comparison methods on both evaluated novelty targets. TC-REJECT provides a practical framework for auditing and conservatively refining partially labeled time series.

1 INTRODUCTION

Labeled time series are used to characterize dynamic systems and train predictive models, yet their annotations are often coarser than the signals themselves. Naturalistic neural recordings may provide broad summaries of observed behavior (Peterson et al., 2022), while clinical labels may summarize protocol blocks or review of longer intervals. The underlying state can change without a corresponding label change, allowing mislabeled windows to contaminate decoders, distort analyses of dynamics, or conceal clinically relevant events. We call a state *novel* when it is absent from the available label vocabulary and *concealed* when it occurs within an interval whose available label does not change. Figure 1A illustrates this problem. A useful label audit must therefore ask both whether a window agrees with its assigned label and whether any available label explains it adequately.

Open-set recognition and out-of-distribution (OOD) detection partly address this problem by identifying observations not represented during training (Geng et al., 2020; Salehi et al., 2022). Common methods use low predictive confidence (Hendrycks & Gimpel, 2016), distance from known classes (Lee et al., 2018), or an explicit unknown-class score (Bendale & Boult, 2016). These approaches are effective when an unseen state is globally unusual, but a mislabeled window may look ordinary relative to the dataset as a whole. Indeed, likelihood and distance can assign unexpectedly high typicality to data from an unknown distribution (Nalisnick et al., 2019; Ren et al., 2021). We therefore distinguish global atypicality from agreement with the available label; the two questions are related, but neither subsumes the other.

We study this distinction under known ground truth using controlled changes in temporal structure and statistical scale. The resulting benchmark shows that global atypicality and label agreement can fail in opposite conditions: a coherent lower-variance state can appear unusually typical even when it conflicts with its label, whereas large scale shifts make global rejection more informative. We

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.

call this observed reversal a *double dissociation*. It is a controlled finding rather than an exhaustive taxonomy of real-world novelty, but it motivates retaining both sources of evidence.

This finding motivates **Typicality-Corrected Rejection (TC-REJECT)**, which compares complementary descriptions of each multichannel window with the known classes and combines their evidence without using examples of the target novel state. Its outputs can flag windows for review, exclusion, or conservative relabeling. Figure 1B summarizes the pipeline.

Our contributions are fourfold. First, we isolate when global atypicality and label agreement succeed or fail under controlled temporal and scale changes. Second, we introduce TC-REJECT, a five-channel detector built from four signal representations, with two fusion routes that require no labeled target novelty. Third, we evaluate withheld and concealed states across synthetic, wearable, ECoG, and DBS data, including a prespecified unseen-session test. Fourth, we assess whether selective label refinement improves downstream classification and report where it does not.

2 Related Work

Open-set, anomaly, and OOD detection. Open-set recognition, anomaly detection, novelty detection, and out-of-distribution (OOD) detection all ask whether an observation is adequately represented by the training data, although they differ in what counts as unknown and what the model should return (Geng et al., 2020; Salehi et al., 2022). Common approaches use classifier confidence (Hendrycks & Gimpel, 2016; Liang et al., 2018), energy (Liu et al., 2020), distance from known classes (Lee et al., 2018; Sun et al., 2022), explicit open-set scores such as OpenMax (Bendale & Boult, 2016), random-partition methods such as Isolation Forest (Liu et al., 2008), or one-class objectives such as One-Class SVM and Deep SVDD (Schölkopf et al., 2001; Ruff et al., 2018). OpenOOD further distinguishes near-OOD sets dominated by semantic shift from far-OOD sets that also introduce an obvious covariate shift (Yang et al., 2022). These methods principally ask whether an observation is unusual relative to the training distribution; auditing an existing label additionally requires asking whether that label adequately describes the observation.

High probability under a model is not equivalent to membership in the training distribution: deep generative models can assign greater likelihood to OOD images than to in-distribution images (Nalisnick et al., 2019), and class-distance scores can degrade for near-OOD examples (Ren et al., 2021). Our study brings this distinction to labeled time series by examining when overall atypicality and agreement with the available label provide different evidence about a hidden state.

Temporal state discovery. Sequence models offer another way to identify temporal changes. The sticky HDP-HMM discovers persistent recurring states without fixing their effective number in advance (Fox et al., 2011), whereas Bayesian online changepoint detection identifies changes in the data-generating process without assigning each segment to a recurring vocabulary (Adams & MacKay, 2007). These methods seek latent temporal structure; our setting instead starts from an available label vocabulary and asks where that vocabulary fails to explain the observed signal.

Conformal evidence and fusion. Conformal prediction maps nonconformity scores to calibrated evidence under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2023), while e-values provide a way to combine valid evidence without assuming independence (Vovk & Wang, 2021). Related work has used labeled outliers to select or combine conformal detectors (Liang et al., 2024), and extensions beyond exchangeability clarify how grouping and distribution shift can weaken nominal guarantees (Barber et al., 2023). These results motivate placing heterogeneous time-series scores on a common empirical scale; Section 3.1 describes the fusion rules and their limitations.

Time-series representations and activity benchmarks. Learned time-series OOD representations such as HyperTF combine temporal and frequency-domain information in supervised, class-conditional embeddings (Lunardi et al., 2026). Our learned representation instead contributes self-supervised evidence alongside covariance, autocorrelation, spectral, and discriminative summaries. We evaluate this broader fusion setting on AJILE12 naturalistic ECoG (Peterson et al., 2022), PAMAP2 wearable activity data (Reiss & Stricker, 2012), and clinical LFP recordings from deep-brain nuclei in two participants who had undergone DBS procedures.

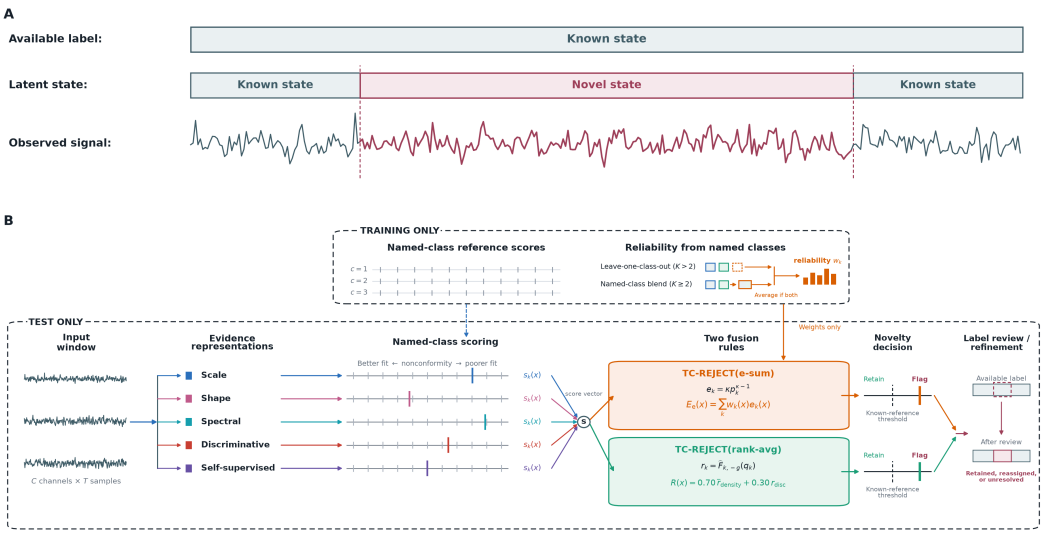

Figure 1: Problem setting and TC-REJECT. **(A)** A coarse label can conceal a state change. **(B)** Four signal representations produce five evidence channels, which are combined by e-value summation or rank averaging to support window-level review or conservative relabeling.

3 METHOD

3.1 TC-REJECT: CALIBRATED FUSION OF COMPLEMENTARY NOVELTY EVIDENCE

TC-REJECT assigns a novelty score to a multichannel window by asking whether any known training class explains that window through five evidence channels derived from four signal representations. Figure 1B gives the order of operations: construct the representations, compare their evidence with the known classes, place the resulting scores on comparable empirical scales, and combine them. The e-sum route preserves the magnitude of the calibrated evidence; the rank-average route uses relative order and is therefore less sensitive to large numerical differences between channels. Both routes use the same channel bank and are reported separately.

Evidence channels. Let $x \in \mathbb{R}^{T\times C}$ denote a window with T samples and C measurement channels. SCALE retains absolute amplitude and covariance. SHAPE normalizes away overall amplitude and captures temporal and cross-channel relationships. SPECTRAL measures class-conditional distance using frequency-band power. The same spectral features also supply the DISCRIMINATIVE channel, which measures classifier support for each known class. LEARNED uses a one-dimensional CNN trained by self-supervised contrastive learning within each dataset and fold; its embedding is compared with the known classes, without cross-dataset pretraining. Thus, four representations produce five evidence channels. Window length, sampling rate, and grouping follow each measurement protocol; Appendix F gives the exact features and training parameters.

Named-class scoring and empirical calibration. Each evidence channel first produces a score measuring how poorly the window fits a known class. In conformal terminology, this is a *nonconformity score*: larger values mean poorer agreement with the reference class (Vovk et al., 2005; Angelopoulos & Bates, 2023). For representation k and named class c, let $z_k(x)$ be the representation of x. The four distance-based channels use

$$s_{k,c}(x) = \frac{1}{2}\big(z_k(x) - \mu_{k,c}\big)^\top P_k\big(z_k(x) - \mu_{k,c}\big),$$

where P_k is a shrinkage estimate of the shared precision matrix. The discriminative channel uses

$$s_{\mathrm{disc},c}(x) = 1 - \widehat{\Pr}(Y = c \mid z_{\mathrm{spectral}}(x)).$$

Each channel then selects its best-fitting named class,

$$c_k^*(x) = \arg\min_c s_{k,c}(x).$$

Thus, the main novelty score asks whether *any* named class explains the window; it does not force the comparison to use the window's available annotation. Own-label compatibility is computed separately when an application performs class-conditional purification or relabeling.

Let $\mathcal{C}_{k,c}$ be the calibration scores of named training windows from class c in the relevant fold. The empirical conformal value is

$$p_k(x) = \frac{1 + \sum_{r \in \mathcal{C}_{k,c_k^*(x)}} \mathbb{I}\left[r \geq s_{k,c_k^*(x)}(x)\right]}{1 + |\mathcal{C}_{k,c_k^*(x)}|}.$$

A small $p_k(x)$ means that even the best-fitting named class provides a poor explanation. Class models, calibration distributions, and the learned encoder exclude the held-out recording group. Because overlapping grouped windows need not be exchangeable, we treat p_k as empirically calibrated and evaluate false-reject behavior rather than assuming exact nominal validity. Run-level standardization, reliability estimation, and adaptive scale weighting use the scored recording without target-novelty labels, making the analysis transductive; Appendix F gives the exact exclusions and aggregation rules.

TC-REJECT(e-sum). The first route converts each p_k to an e-value using the Vovk–Wang calibrator

$$e_k(x) = \kappa p_k(x)^{\kappa - 1}, \qquad \kappa = 0.4.$$

For valid inputs, a fixed nonnegative weighted average remains an e-value without requiring channel independence (Vovk & Wang, 2021). Base weights are estimated from known-class holdout and mixture probes, without labels from the target novel state. The reported score also uses a window-dependent gate to suppress confidently inverted scale evidence. Because that gate is data dependent, the adaptive score does not inherit the fixed-weight guarantee; its threshold is evaluated by leave-one-group-out calibration. Appendix F gives the probe, weighting, and gate definitions.

TC-REJECT(rank-avg). The second route reduces sensitivity to channel magnitude and probe weights. It orients every score so that larger values indicate novelty, maps each channel to its leave-one-group-out reference rank, and averages the four density-family ranks into $D(x)$. The final score is

$$S_{\text{rank}}(x) = 0.70 D(x) + 0.30 R_{\text{disc}}(x).$$

The discriminative weight is 0.30, as prespecified for the 12 previously unused AJILE12 sessions based on the AJILE12 and PAMAP2 analyses. Rank averaging does not use the e-value threshold; for either route, operational decisions use the $(1 - \alpha)$ leave-one-group-out quantile of known-reference scores, with $\alpha = 0.10$ unless otherwise specified. AUROC uses continuous scores. Channel orientation and rank definitions appear in Appendix F.

3.2 Controlled benchmark for concealed temporal novelty

We construct a controlled benchmark in which a multiregion linear dynamical system generates 12-channel recordings with two named states. A novel state is inserted within a labeled block without changing the available annotation, so the detector sees the signal and the coarse block label but not the concealed transition. Figure 2A illustrates three conditions: a structured alternative process with lower realized variance than the named states, a large unstructured input that inflates variance, and rapidly changing dynamics that alter both temporal structure and scale.

Before fusing channels, we compare the two questions that motivate TC-REJECT. Density-based rejection asks whether a window lies far from every known state, using minimum class-conditional Mahalanobis nonconformity. Label-relative evidence asks whether the window agrees with its assigned block label, using $1 - \widehat{\Pr}(Y = y_{\text{avail}} \mid x)$. We then vary the novel-to-named variance ratio while holding the alternative dynamics fixed and use complementary sweeps to distinguish changes in scale from changes in dynamical shape. Fitting and evaluation are block-disjoint, and the concealed state is used only to define evaluation targets. Appendix A gives the generator, variance-ratio definition, sweeps, windowing, and evaluation protocol; Appendix B describes the temporal-state comparator.

3.3 MEASURED-DATA EVALUATION PROTOCOLS

AJILE12. We use AJILE12, a dataset of long-duration naturalistic ECoG recordings with synchronized behavioral annotations from participants with clinically implanted electrodes (Peterson et al., 2022). We evaluate structural and blend-type novelty in one prespecified session from each of ten participants, using contiguous recording blocks to define out-of-fold scoring. Structural targets are behaviorally annotated states outside the named vocabulary; the blend-type target is the unresolved subset of compound annotations after named-class fit refinement. Detailed label construction, channel construction, windowing, and downstream-classification procedures are given in Appendix C.

PAMAP2. We use PAMAP2, a benchmark dataset of wearable inertial and physiological measurements recorded during protocol activities (Reiss & Stricker, 2012). We withhold five activities across three definitions and evaluate by held-out participant using 10-second windows. No withheld-activity label supervises fold-specific model fitting; the run-level transductive preprocessing described in Section 3.1 remains in effect. The complete split definitions and baseline implementations are given in Appendix D.

DBS-LFP. We use two non-intraoperative LFP recordings collected under a previously described staged-DBS protocol (Liker & Sanger, 2019; Liker et al., 2024). Each analyzed session contains four reaching-movement blocks interleaved with six Rest blocks, and independent review identified additional movements within clinically labeled Rest. Clinical segments define the cross-validation folds. Detection is evaluated across all scored windows, whereas relabeling is evaluated only among windows originally labeled Rest. Appendix E provides participant, acquisition, ethics, windowing, and decision-rule details.

4 RESULTS

4.1 SYNTHETIC BENCHMARK: COMPLEMENTARY DETECTION ACROSS NOVELTY CONDITIONS

Global atypicality and label agreement succeed in different conditions. Figure 2B shows that the global outlier score ranks the coherent lower-scale novel state in the wrong direction, whereas checking agreement with the assigned label remains informative. Their relative strengths reverse when novelty produces a large scale increase. This observed double dissociation shows that neither question is sufficient by itself; detailed condition-wise scores are reported in Appendix B.1.

Relative scale determines the rejection crossover. Changing the novel-state variance while holding its dynamics fixed moves the global outlier score monotonically through chance (Figure 2C). Changing dynamical shape at approximately fixed variance does not reproduce this pattern. Within this benchmark, relative statistical scale therefore explains the score inversion more directly than dynamical difference alone; sweep details are given in Appendix B.1.

TC-REJECT maintains detection across the three conditions. Across ten independently generated datasets, e-value-sum fusion reaches median AUROC of 0.816, 0.987, and 0.993 for structured lower-scale, scale-inflated, and scale-divergent novelty, respectively (Figure 2D), compared with 0.678, 0.883, and 0.981 for rank averaging. Thus, combining complementary representations preserves useful detection across conditions in which the underlying evidence sources have opposing strengths. Additional summaries and temporal-model comparisons appear in Appendices B.1 and B.2.

4.2 AJILE12: NOVELTY WITHIN COARSE BEHAVIORAL ANNOTATIONS

AJILE12 contains long-duration naturalistic ECoG recordings with temporally coarse behavioral annotations (Peterson et al., 2022). We evaluate structural novelty and unresolved windows from compound annotations in one prespecified session from each of ten participants, using Talk and Computer/phone as the shared named vocabulary. Appendix C gives the complete target construction, windowing, and preprocessing.

Figure 3A illustrates window-level refinement of compound blocks: windows supporting an available named component receive that label, windows supporting a different named class are withheld as

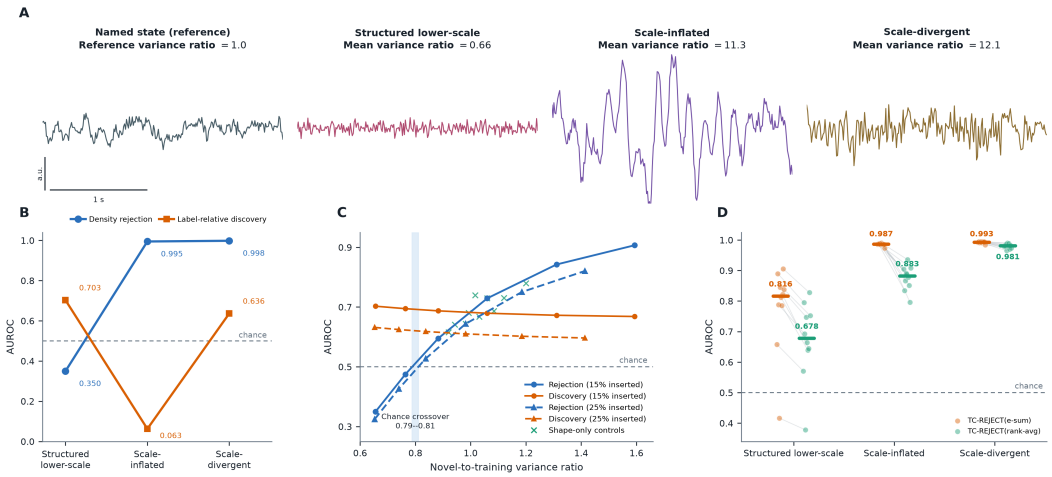

Figure 2: Synthetic benchmark. **(A)** A known state and three inserted novelty conditions. **(B)** AUROC of global atypicality and label agreement. **(C)** Effects of relative variance and dynamical shape. **(D)** AUROC of the two fusion rules across ten independently generated datasets; thick bars denote medians.

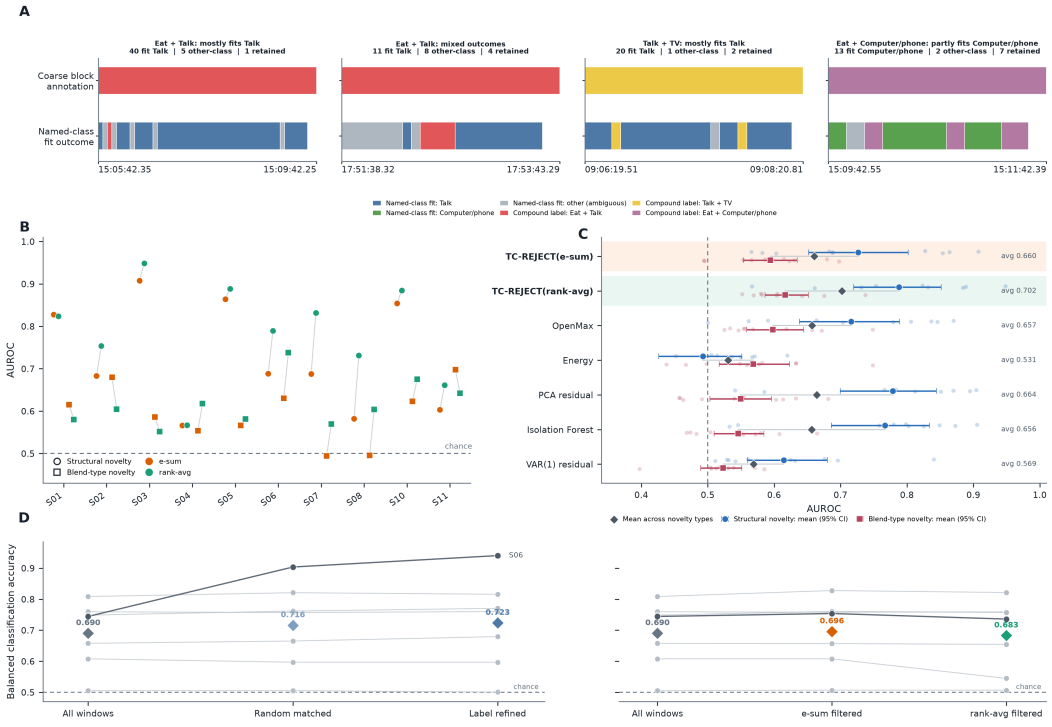

Figure 3: AJILE12 evaluation. **(A)** Compound annotations and 5-second-hop refinement; gray denotes ambiguous support. Because 10-second windows overlap, decision strips end one hop before complete blocks. **(B)** Paired fusion AUROCs for one session per participant and target. **(C)** Independent baselines; points denote participants, intervals bootstrap 95% confidence intervals, and diamonds target-averaged means. **(D)** Downstream Talk-versus-Computer/phone classification performance, measured by balanced accuracy, after label refinement or score filtering; points denote participants and diamonds denote means.

ambiguous, and unsupported windows remain novel. These revisions are model-derived rather than frame-accurate behavioral ground truth.

Rank-average fusion improves the cohort mean for both targets (Figure 3B), reaching AUROC 0.788 for structural novelty and 0.617 for the unresolved blend-type target, compared with 0.726 and 0.594 for e-value summation. Rank-average TC-REJECT also has the highest mean across the two targets among the independent methods in Figure 3C (0.702), although participant-level rankings vary. Because the blend target is derived from the named-class fit rule rather than independent review, its AUROC is descriptive. Participant-level results are reported in Appendix G.

We then evaluated the prespecified fusion rules on 12 previously unused sessions under a prespecified protocol. Rank averaging improves the prespecified session-level metric by 0.086 AUROC over e-value summation, is higher in 11 of 12 sessions, and is highest on both targets among the independent methods tested. The paired subject-bootstrap 95% confidence interval is $[0.049, 0.123]$; the full baseline comparison appears in Appendix H.

Figure 3D tests downstream Talk-versus-Computer/phone classification using electrode-wise XGBoost models on gamma-band features and majority voting. Mean balanced accuracy is 0.690 with all windows, 0.716 after size-matched random removal, and 0.723 after label refinement; e-value and rank-score filtering reach 0.696 and 0.683. Refinement is numerically highest, but its small margin over random removal indicates that data reduction explains part of the gain. This transductive analysis is descriptive; complete procedures and participant results appear in Appendices C and K.

4.3 PAMAP2: GENERALIZATION TO HELD-OUT WEARABLE ACTIVITIES

PAMAP2 records 12 protocol activities from nine participants using wearable inertial sensors at the hand, chest, and ankle (Reiss & Stricker, 2012). We evaluate five withheld activities—vacuum cleaning, rope jumping, Nordic walking, and ascending and descending stairs—while holding out participants. The remaining activities form the named vocabulary. Appendix D gives the complete splits and preprocessing.

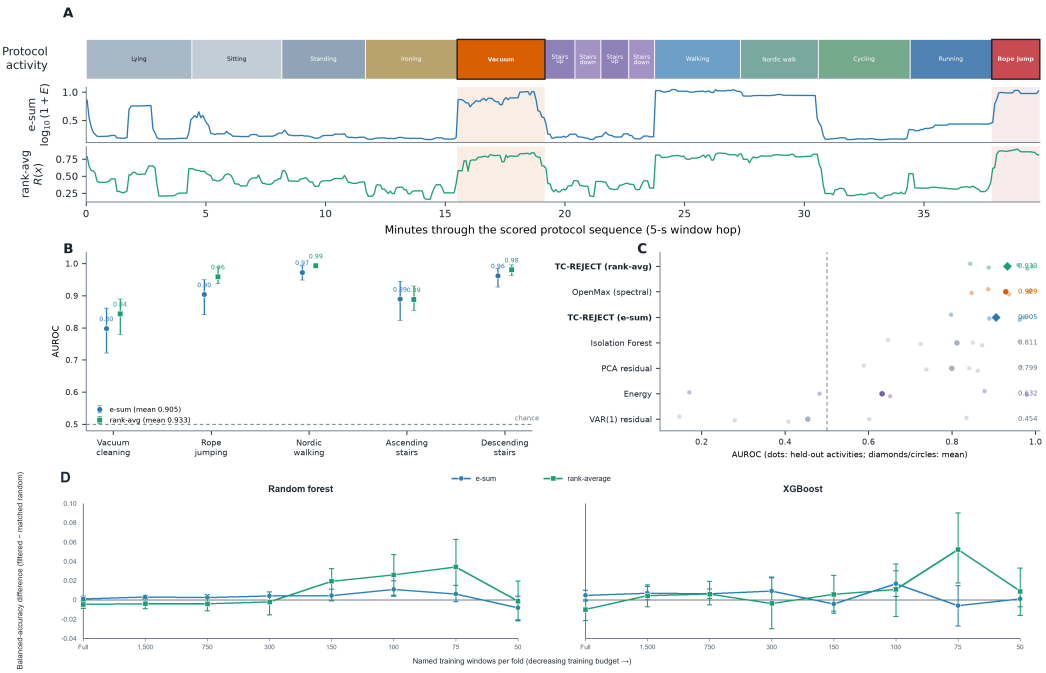

Figure 4: PAMAP2 results. **(A)** Activity timeline and novelty scores. **(B)** AUROC by withheld activity. **(C)** Mean performance across activities (faint points). **(D)** Paired ten-class balanced-accuracy difference between score-informed and matched-random removal, by budget and classifier (participant-bootstrap 95% CIs).

Figure 4A shows that both fusion scores rise over withheld activities, while occasional high scores on named activities expose the remaining false positives. Across all five held-out activities (Figure 4B–

C), rank-average fusion reaches mean AUROC 0.933, compared with 0.905 for e-value summation, and is higher for four activities. It has the highest displayed mean, narrowly above the OpenMax-style spectral baseline at 0.929; the ordering varies by activity, so this difference does not establish uniform superiority. Full activity-wise estimates and the baseline panel appear in Appendix I.

Using the finalized scores at $\alpha = 0.10$ over five seeds, rank-average filtering outperforms size-matched random removal at 75 named windows per fold with both random forests ($\Delta = 0.034$, 95% CI $[0.006, 0.063]$) and XGBoost ($\Delta = 0.052$, $[0.018, 0.090]$). At 100 windows the gain remains positive for random forests but is inconclusive for XGBoost; it is absent at budgets of at least 150 windows and at 50 windows. E-value-sum effects are smaller and inconsistent across budgets. Thus, informed trimming can help in a limited scarcity regime but does not provide a general downstream improvement. Full paired results and the classifier cross-check appear in Appendix K.

4.4 DBS: DETECTING AND RELABELING EXTRA MOVEMENT INSIDE CLINICAL REST

We evaluate TC-REJECT on clinical LFP recordings from deep-brain nuclei in two participants who had undergone DBS procedures. Independent review identifies additional movements within clinically annotated Rest intervals, providing a measured test of concealed label contamination. Appendix E gives the acquisition, windowing, and evaluation details.

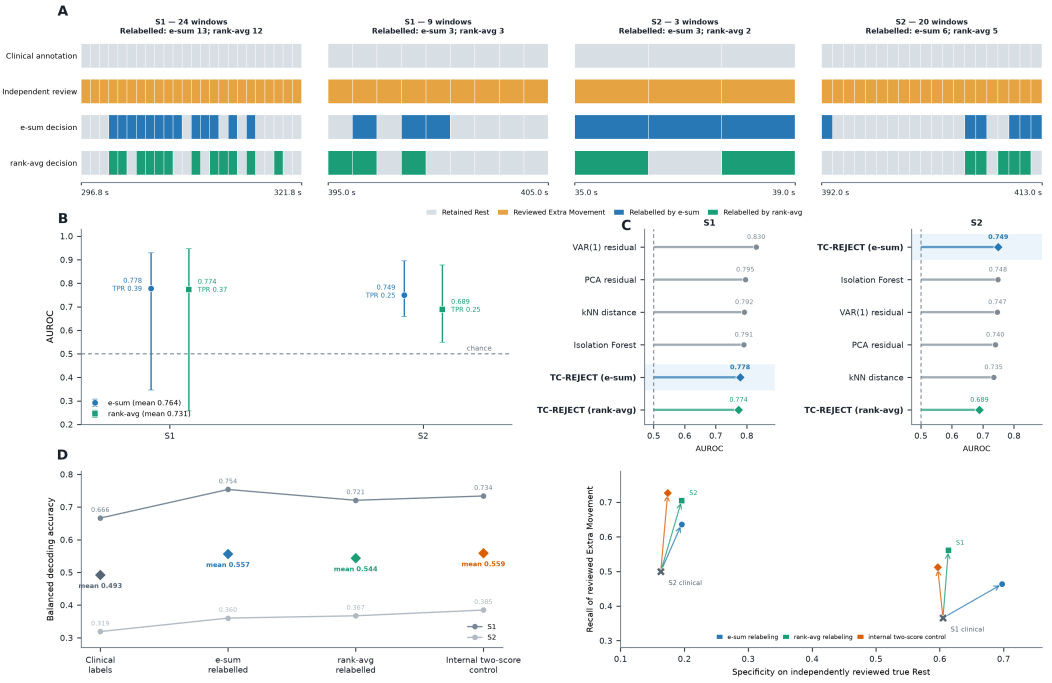

Figure 5: DBS evaluation. **(A)** Clinical Rest labels, independent movement review, and TC-REJECT reassignments. **(B)** Extra-movement detection AUROC with segment-bootstrap intervals. **(C)** Detector comparisons by participant. **(D)** Post-hoc Rest-versus-Movement balanced accuracy after relabeling; the two-score condition averages spectral Mahalanobis and maximum-softmax ranks.

Within clinically annotated Rest, e-value-sum fusion identifies a high-confidence subset of the independently reviewed movements (Figure 5A): precision is 0.941 and 0.647, while recall is 0.390 and 0.250 for S1 and S2, respectively. The limited recall means that the revised labels are selective rather than complete.

Detection performance and method ranking differ between participants (Figure 5B–C). E-value-sum fusion reaches AUROC 0.778 on S1 and 0.749 on S2; several baselines are higher on S1, whereas e-value-sum TC-REJECT is narrowly highest among the displayed methods on S2. With only four and five independently reviewed Extra-Movement segments, the confidence intervals overlap broadly and do not identify a best detector.

In the post-hoc downstream analysis, e-value-sum relabeling raises Rest-versus-Movement balanced accuracy from 0.666 to 0.754 in S1 and from 0.319 to 0.360 in S2 (Figure 5D). This consistent direction is encouraging, but the two-recording cohort supports only a descriptive conclusion. Full relabeling, baseline, and decoder results appear in Appendix J.

5 DISCUSSION

Partially labeled time series create a failure mode that conventional OOD detection does not fully address: a window can be typical of the overall dataset while no longer matching the label assigned to its interval. A coherent lower-variance state appears unusually typical despite conflicting with its block label, whereas large scale shifts make global rejection informative while label agreement can fail. This *double dissociation* is the central conceptual contribution of the work: auditing labels requires both global atypicality and label-relative evidence. It connects unexpected OOD likelihood and distance behavior (Nalisnick et al., 2019; Ren et al., 2021) to a concrete failure mode in labeled time series and explains why a single novelty score can be unreliable.

TC-REJECT turns this insight into a practical auditing framework. It integrates amplitude and covariance, temporal shape, frequency content, known-class support, and self-supervised temporal structure, then combines these complementary views through either evidence-preserving e-value fusion or scale-robust rank fusion. Without labeled examples of the target novel state, TC-REJECT can flag windows for expert review, exclusion from training, reassignment to a supported known class, or retention as unresolved novelty.

The empirical results show that this framework transfers across distinct signals and novelty definitions. PAMAP2 tests activities withheld from the training vocabulary, whereas AJILE12 and DBS test the more difficult setting in which a state is concealed inside an existing annotation. Rank-average TC-REJECT achieves the highest mean among the independent methods on both targets in the prespecified 12-session AJILE12 evaluation and the highest five-activity mean on PAMAP2, narrowly ahead of OpenMax. The controlled benchmark favors e-value fusion, while the independent AJILE12 sessions favor rank fusion, illustrating why reporting both routes is informative. The small DBS cohort does not establish a best detector, but the precision-oriented relabeling results show that TC-REJECT can identify a conservative subset of clinically mislabeled windows.

This capability is particularly relevant to brain-computer interfaces and other supervised learning systems built from physiological time series. BCI decoders learn mappings between neural activity and behavioral, cognitive, or clinical labels; when those labels are coarse, delayed, or internally inconsistent, the resulting models can reproduce annotation errors rather than the intended state. Because neural recordings are expensive to acquire and labels are often reused across analyses, an undetected labeling error can propagate through feature discovery, decoder training, benchmarking, and scientific interpretation. TC-REJECT provides an upstream quality-control layer that can prioritize scarce expert attention and expose training windows that deserve removal or relabeling before they influence a downstream model. Better auditing cannot guarantee better prediction, but it can improve the reliability, interpretability, and reproducibility of the data on which those models depend.

The results also set clear limits. The synthetic study covers three deliberately constructed conditions rather than every possible temporal change. AJILE12 lacks frame-accurate behavioral ground truth, its blend-type target is partly model derived, and the DBS study contains only two recordings with four to five independently reviewed Extra-Movement segments per participant. Downstream benefits are correspondingly mixed: the AJILE12 gain beyond matched random removal is modest, PAMAP2 shows a bounded benefit only under moderate scarcity, and the DBS decoder analysis is post hoc. The PAMAP2 pattern is limited to one split, one pinned score run, five seeds, eight participants, and two adjacent budgets, so it does not establish a broad learning curve. The implementation also uses information from the scored run for standardization and adaptive weighting, and empirical calibration remains imperfect for grouped, overlapping windows; nominal thresholds should not be interpreted as deployment guarantees (Barber et al., 2023). Future work should test broader dynamics, develop fully inductive group-aware calibration, incorporate complementary temporal models, and evaluate prespecified relabeling rules prospectively in larger clinical and BCI cohorts. TC-REJECT should therefore be used as a label-review and decision-support layer, not as an autonomous source of ground truth.

AI USE STATEMENT

Under detailed author direction, generative AI tools assisted with manuscript drafting and editing, literature organization, figure prototyping, and portions of analysis-code implementation and revision. The authors designed the study, conducted and directed the analyses, interpreted the results, and made all scientific decisions. All numerical results were produced by executing the analysis code and verified against the underlying outputs. The authors reviewed and revised all AI-assisted material and take responsibility for the submission.

ETHICS STATEMENT

AJILE12 and PAMAP2 are publicly available datasets and were analyzed in accordance with their data-use conditions. The DBS-LFP analysis uses human electrophysiological recordings collected under protocols approved by institutional review boards. Prior to surgery, written informed consent and HIPAA authorization for research use were obtained from participants or their legal guardians, as applicable; the participating institution names are withheld for double-blind review. Clinical recordings are not publicly redistributed because of privacy protections. TC-REJECT is intended to support research and expert label review, not autonomous clinical diagnosis or treatment.

REPRODUCIBILITY STATEMENT

Code for TC-REJECT, synthetic-data generation, experiment configurations, and scripts used to produce the reported tables and figures are available at `https://github.com/anonymousiclr01`. The repository documents dataset preparation, evaluation splits, random seeds, and commands for reproducing the analyses. Public datasets should be obtained from their cited sources. The clinical DBS recordings contain protected health information and are not publicly redistributed; de-identified data may be available on reasonable request, subject to participant-consent and institutional data-access requirements.

REFERENCES

Ryan Prescott Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1563–1572, 2016.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Mario A. T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, 24(3):381–396, 2002.

Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, pp. 1020–1056, 2011.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Ziyi Liang, Matteo Sesia, and Wenguang Sun. Integrative conformal p-values for out-of-distribution testing with labelled outliers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):671–693, 2024.

M Liker and T Sanger. Pediatric deep brain stimulation in secondary dystonia using stereotactic depth electrode targeting: 1304. *Movement Disorders*, 34:S534–S535, 2019.

Mark A Liker, Terence D Sanger, Jennifer A MacLean, Jaya Nataraj, Enrique Arguelles, Mark Krieger, Aaron Robison, and Joffre Olaya. Stereotactic awake basal ganglia electrophysiological recording and stimulation (sabers): A novel staged procedure for personalized targeting of deep brain stimulation in pediatric movement and neuropsychiatric disorders. *Journal of child neurology*, 39(1-2):33–44, 2024.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pp. 413–422. Ieee, 2008.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Willian T. Lunardi, Samridha Shrestha, and Martin Andreoni. Learning hyperspherical time–frequency representations for time-series out-of-distribution detection. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 2888–2896, 2026. doi: 10.24963/ijcai.2026/321.

Conor McCarthy, Loes Quirijnen, Jan Peter van Zandwijk, Zeno Geradts, and Marcel Worring. Hi-oscar: Hierarchical open-set classifier for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(4):1–27, 2025.

Maryam Ostadsharif Memar, Navid Ziaei, and Behzad Nazari. A combined channel approach for decoding intracranial eeg signals: Enhancing accuracy through spatial information integration. In *2025 33rd International Conference on Electrical Engineering (ICEE)*, pp. 363–368. IEEE, 2025.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1xwNhCcYm`.

Steven M. Peterson, Satpreet H. Singh, Benjamin Dichter, Michael Scheid, Rajesh P. N. Rao, and Bingni W. Brunton. AJILE12: Long-term naturalistic human intracranial neural recordings and pose. *Scientific Data*, 9:184, 2022. doi: 10.1038/s41597-022-01280-y.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pp. 108–109. IEEE, 2012.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *Transactions on Machine Learning Research*, 2022.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pp. 20827–20840. PMLR, 2022.

Vladimir Vovk and Ruodu Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

A Synthetic benchmark construction details

Recordings and available labels. Each synthetic dataset contains eight recordings sampled at 100 Hz. A recording contains eight consecutive blocks whose durations are drawn between 0.85 and 1.15 times 30 seconds. The available block labels alternate between two named states. Each state is generated by a stable four-dimensional linear dynamical system. Four simulated regions share part of their dynamics but retain region-specific structure; an independent mixing matrix maps each region to three observed channels, producing 12 channels in total. Additive observation noise is set to 10 dB signal-to-noise ratio, and each observed channel is standardized within its recording.

The available annotation remains fixed over an entire block even when its latent state changes. For evaluation of the two TC-REJECT fusion rules, novel states occupy approximately 25% of each block and segments from the other named state occupy approximately 10%; the realized fractions vary by up to 0.08 across blocks. Insertions are divided among one to three nonoverlapping segments with a minimum duration of one second. The additional named-state insertions create closed-set label errors, whereas the three novel states create concealed open-set errors. The detector receives only the block label. The sample-level latent state is retained solely to construct evaluation targets.

Novel-state constructions. The three inserted novel states are designed to separate a coherent change in dynamics from a large change in signal scale. The structured lower-scale state replaces the named process with a distinct, persistent transition matrix. The scale-inflated state retains the enclosing named transition dynamics but adds strong unstructured innovations. The scale-divergent state uses a new transition matrix every 0.5 seconds and therefore lacks a persistent temporal template. Table A.1 reports their measured variance relative to the average named-state variance in the controlled comparison.

Table A.1: Synthetic novelty constructions and baseline variance ratios.

Term used here	Construction	Baseline variance ratio
Structured lower-scale novelty	A stable alternative linear dynamical process replaces part of a named block without added high-amplitude drive. Its realized variance is lower than the average named-state variance.	≈ 0.66
Scale-inflated novelty	Additional unstructured drive is superimposed on the enclosing named process. The original dynamics remain present, but total variance increases.	≈ 11.3
Scale-divergent novelty	The dynamics are redrawn every 0.5 seconds, eliminating a persistent template and producing a large departure in scale.	≈ 12.1

Variance-ratio intervention. To separate marginal scale from dynamical structure, we quantify the variance of latent state z relative to the named-state mixture as

$$\rho_v(z) = \frac{C^{-1}\sum_{j=1}^{C}\mathrm{Var}(X_j \mid Z = z)}{|\mathcal{Y}|^{-1}\sum_{c\in\mathcal{Y}} C^{-1}\sum_{j=1}^{C}\mathrm{Var}(X_j \mid Y = c)},$$

where $\mathcal{Y}$ is the set of named states and C is the number of observed channels. We vary the innovation amplitude of the stable alternative process while holding its transition dynamics fixed, thereby changing ρ_v without changing the identity of that process. A spectral-radius sweep changes persistence and variance together, whereas a rotation-angle sweep changes dynamical shape at approximately matched variance.

Windows and evaluation groups. The observations are divided into 0.5-second windows with a 0.2-second stride. Each window inherits the available label of its enclosing block, while its evaluation target is determined by the latent state at the window center. Blocks define the grouping variable and are separated across five folds. The controlled evidence comparison in Figure 2B uses a 0.15 inserted-novelty rate; the five-channel fusion evaluation uses 0.25. Both use 0.10 closed-set contamination. Results are reported separately for the three novel states rather than pooling them into a single target.

B SYNTHETIC TEMPORAL-MODEL COMPARATOR

The temporal comparator is a hidden Markov model containing named states and candidate novel states. Its emission distribution is a tied factor-analysis Gaussian mixture,

$$\Sigma = \mathrm{diag}(\psi) + WW^\top,$$

and it exposes separate REJECT and DISCOVER readouts that operate in parallel.

REJECT (p_{gate}). REJECT measures absolute atypicality. It ranks a window's log density against calibration windows treated as named and forms a likelihood ratio between a $\mathrm{Beta}(1, \beta)$ novel model and the uniform named distribution induced by conformal calibration. Because inserted novel intervals retain their block labels, the reference distribution is trimmed according to the specified contamination rate.

DISCOVER (p_{tmpl} and p_{disagree}). The template readout assigns posterior mass to candidate novel states that compete directly with the named components. The second readout is label-relative,

$$p_{\mathrm{disagree}}(x) = 1 - p(y_{\mathrm{avail}} \mid x),$$

and asks whether the observed window supports its available label. The analogous real-data quantity is the own-label-compatibility score used for purification and relabeling (Section 3.1); it is kept separate from the main any-class novelty score.

We allocate six candidate discovery states and use an entropy-based occupancy penalty inspired by component-pruning approaches for over-specified finite mixtures (Figueiredo & Jain, 2002); the template and density readouts are combined by their maximum. The factor-analysis covariance retains cross-feature dependence while controlling the number of covariance parameters, and the generative fit uses a trimmed likelihood so that concealed novel windows do not dominate estimation of the named-state density.

B.1 SYNTHETIC EVALUATION SUMMARY

We evaluate the two reported fusion rules with the same five-channel architecture on ten independently generated datasets. Table B.1 reports median AUROC for each novelty condition.

Table B.1: Median AUROC for the two TC-REJECT fusion rules on ten independently generated synthetic datasets.

Fusion rule	Datasets	Structured lower-scale	Scale-inflated	Scale-divergent
E-value sum	10	**0.816**	**0.987**	**0.993**
Rank average	10	0.678	0.883	0.981

E-value-sum fusion is higher in all three controlled conditions, with the largest separation under structured lower-scale and scale-inflated novelty. Both rules remain strongest for scale-divergent novelty, whereas structured lower-scale novelty is the most variable condition. Figure 2D displays the paired results for all ten datasets.

Before fusion, the global outlier and label-agreement scores show complementary behavior under block-disjoint cross-validation. Across three independently generated datasets, their mean AUROCs are 0.350 versus 0.703 for structured lower-scale novelty, 0.995 versus 0.063 for scale-inflated novelty, and 0.998 versus 0.636 for scale-divergent novelty. In controlled variance sweeps, the global outlier score crosses chance at novel-to-named variance ratios of approximately 0.74–0.86. Rotation-angle changes at approximately fixed variance do not reproduce this crossover.

B.2 Temporal-model comparator results

Table B.2 reports the performance of the temporal-state comparator described in Appendix B on ten independently generated datasets with an inserted-novelty rate of 0.25. The conformal reference is trimmed according to that rate because concealed novel intervals retain their available block labels.

Table B.2: Median AUROC of the temporal-model readouts. Density readouts use ten synthetic datasets; the structured lower-scale template cell uses nine because one seed had no live template slot.

Novelty condition	Evaluated readout	Median AUROC
Structured lower-scale	DISCOVER (template; $n = 9$)	0.474
Scale-inflated	REJECT (density)	0.941
Scale-divergent	REJECT (density)	0.964

The density readout detects both scale-shifted conditions, but the template readout remains near chance for structured lower-scale novelty. The trained temporal model therefore supports the density-rejection result but does not reproduce the label-relative discovery result. The double-dissociation claim is consequently based on the block-disjoint controlled comparison in Figure 2B rather than on this temporal comparator. Relative to the closed-set control, named-state accuracy decreases by a median of 0.41 percentage points.

C AJILE12 methods

AJILE12 contains long-duration naturalistic ECoG recordings and synchronized behavioral annotations from 12 participants with clinically placed electrodes (Peterson et al., 2022). Ten participants have sufficient support for the common cohort evaluation. Talk and Computer/phone form the named vocabulary. Sleep/rest, TV, Inactive, and Other activity form the structural-novelty target. Compound annotations such as Talk+TV, Eat+Talk, and Eat+Computer/phone provide candidate blend-type windows. A scale-based named-class fit rule reassigns windows that support an available named component, excludes windows that instead support a different named class as ambiguous, and retains the unresolved subset as the blend-type target. Structural labels are independent coarse behavioral annotations, whereas the blend-type target is partly model derived and should not be interpreted as frame-accurate ground truth.

Signals are divided into 10-second windows with a 5-second stride. Contiguous annotation blocks define the folds, and fold-specific class models, learned encoders, and calibration distributions exclude the target block. Run-level hand-crafted-feature standardization and reliability estimation remain transductive as described in Section 3.1. The same five-channel bank—scale, shape, spectral, discriminative, and learned—is computed from the available ECoG channels. Overlapping windows from the same behavioral block are not treated as independent: uncertainty is estimated by resampling contiguous blocks. Structural and blend-type AUROC are calculated separately for each participant and then summarized across participants. Because SCALE defines the refinement that constructs the unresolved blend-type subset, blend-type detector comparisons are descriptive rather than an independent validation against external labels.

For compound-block refinement, a window is reassigned only when its signal supports a named component available in the block annotation. Support for a different named class is treated as ambiguous and excluded from the refined training set; windows unsupported by any named class retain novel status. The same held-out blocks are used for downstream evaluation under the full, refined, and size-matched random-removal training conditions. The downstream classifier adapts the combined-channel intracranial decoder of Memar et al. (2025) to Talk versus Computer/phone: each 10-second parent window is divided into five nonoverlapping 2-second windows, gamma-band

features are classified by a separate XGBoost model (Chen & Guestrin, 2016) for each electrode, and electrode decisions are combined by majority vote after greedy electrode selection on block-disjoint validation data. Within each training fold, SMOTE (Chawla et al., 2002) balances the minority class before the electrode-specific models are fitted.

D PAMAP2 PROTOCOL DETAILS

PAMAP2 records 12 protocol activities from nine participants using 100-Hz triaxial accelerometers at the hand, chest, and ankle (Reiss & Stricker, 2012). Consistent with leave-class-out evaluations in open-set activity recognition (McCarthy et al., 2025), we use 10-second windows with 50% overlap and leave-one-participant-out evaluation. Three activity sets are withheld: (A) rope jumping and vacuum cleaning, (B) Nordic walking while ordinary walking remains named, and (C) ascending and descending stairs. These splits span distinct activities and a near-neighbor comparison, and all five withheld activities are reported. Fold-specific class models, learned encoders, and calibration distributions exclude the target participant, whereas run-level standardization, reliability estimation, and adaptive scale weighting are transductive as described in Section 3.1. The downstream ten-class budget analysis includes eight participants because one fold lacks sufficient per-class support; this eligibility criterion is applied before method comparison.

Energy (Liu et al., 2020) and an OpenMax-style score (Bendale & Boult, 2016) are computed from the scale and spectral representations using the same held-out-participant folds.

E DBS-LFP RECORDING DETAILS

We analyze non-intraoperative local field potential (LFP) recordings from two participants, denoted S1 and S2, collected under a staged-DBS protocol and acquisition framework (Liker & Sanger, 2019; Liker et al., 2024). Recordings were obtained one day after implantation of temporary bilateral DBS leads. Lead placement was verified by co-registering preoperative MRI with postoperative CT, and signals were acquired from high-impedance microwire contacts on AdTech MM16C leads at approximately 24 kHz. Table E.1 provides the participant characteristics relevant to this analysis.

Table E.1: Clinical and recording characteristics of the two participants. Abbreviations: CP, cerebral palsy; KMT2B, lysine methyltransferase 2B; GPi, globus pallidus internus; Vo, ventral oralis; STN, subthalamic nucleus; VIM, ventral intermediate nucleus.

ID	Etiology	Dystonia type	Recorded targets
S1	CP	Generalized dystonia	GPi, Vo, STN, VIM
S2	KMT2B	Generalized dystonia	GPi, Vo, STN, VIM

For the session analyzed here, each participant performed four reaching-movement blocks interleaved with six Rest blocks. We retain 40 channels spanning bilateral target structures, resample the signals to 100 Hz, and divide them into 2-second windows with a 1-second stride. The available annotation vocabulary contains Rest and Movement. Independent review additionally identifies Extra Movement intervals embedded within clinically labeled Rest, providing a direct test of concealed label contamination. S1 contains 41 Extra Movement windows among 398 scored windows; S2 contains 44 among 374.

Human-subject protections, consent procedures, and data-access restrictions are summarized in the Ethics Statement.

Clinical segments define the cross-validation folds: each segment is scored using models and thresholds estimated from the remaining segments. Detection compares Extra Movement windows with the named Rest and Movement windows. Relabeling is evaluated only among windows originally labeled Rest, using $\alpha = 0.1$. Method comparisons use OpenMax-style, energy, One-Class SVM (Schölkopf et al., 2001), nearest-neighbor, subspace, Isolation Forest (Liu et al., 2008), and temporal-prediction detectors. Precision is the fraction of reassigned Rest windows independently reviewed as Extra Movement; recall is the fraction of Extra Movement windows recovered; and the false-relabel rate is the fraction of independently reviewed Rest windows incorrectly reassigned.

F SHARED IMPLEMENTATION DETAILS

All hand-crafted representations are standardized once using available named windows from the scored run; this is the run-level transductive preprocessing disclosed in Section 3.1. SCALE and SHAPE include zero-lag covariance or correlation together with lagged statistics at 2, 5, and 10 samples. SPECTRAL uses Welch segments of at most 200 samples with a hop of at most 100 samples and log power in bands with edges at 0, 3, 6, 10, and 15 Hz, clipped at the Nyquist frequency. Feature spaces above 200 dimensions are reduced by PCA within the fitting fold. The class-conditional scaffold uses shrinkage 0.1 and falls back to the pooled calibration distribution when a named class contains fewer than five calibration windows.

The learned encoder has three one-dimensional convolutional layers with 32, 64, and 32 output channels, kernel width 7, stride 2, batch normalization, rectified-linear activations, and adaptive average pooling, followed by a two-layer projection head during contrastive training. It is trained for 20 epochs with Adam (learning rate 10^{-3}), batch size 128, NT-Xent temperature 0.2 (Chen et al., 2020), and a 32-dimensional embedding. Augmentations comprise additive jitter, per-window channel scaling, channel dropout, and mild random cropping followed by temporal resampling. The encoder is retrained in each fold after excluding that fold's held-out groups.

Grouped calibration and transductive preprocessing. Target windows are scored only by folds that exclude their recording group; a target window scored by multiple folds receives the mean of its fold-specific values. For distance-based channels, the calibration window's own group is also excluded from the class mean used for its calibration distance. The class models, calibration distributions, and learned encoder therefore exclude the held-out participant or recording. Available named windows from the scored run nevertheless determine run-level standardization and reliability estimates, and the unlabeled score distribution sets the adaptive scale gate. The target novelty labels are never used for these operations.

E-value weighting and adaptive scale gate. One reliability probe temporarily treats each known class as unknown, and a second forms synthetic mixtures of pairs of known classes. With only two named classes, the leave-one-class-out probe is uninformative and only the mixture probe is used. Otherwise, the two probe AUROCs are averaged. Improvement above chance is normalized and capped so that no base weight exceeds 0.65; if no valid mixture pair exists, base weights are equal.

Let $d_\perp(x)$ be the distance of the scale representation from the affine hull of the named-class centroids, and let a and b be its 25th and 75th percentiles in the unlabeled scored run (with $b = a + 10^{-6}$ when they coincide). The scale gate and fused score are

$$u(x) = \mathrm{clip}\left(\frac{d_\perp(x) - a}{b - a}, 0, 1\right), \qquad w_s(x) = 0.55u(x)^2\big(3 - 2u(x)\big),$$

$$E_{\mathrm{e}}(x) = w_s(x)e_s(x) + (1 - w_s(x)) \sum_{k \neq s} \frac{\bar{w}_k}{\sum_{j \neq s} \bar{w}_j} e_k(x),$$

where $\bar{w}_k$ is the known-class-derived base weight.

Rank-average implementation. Each channel is oriented so that larger values indicate novelty. With two named classes, scale, spectral, and discriminative use their raw Mahalanobis, spectral Mahalanobis, and maximum-softmax rejection scores; with three or more classes, they use $1 - p_k$. Shape and learned always use $1 - p_k$. For oriented score q_k, named reference set $\mathcal{R}$, and recording group g,

$$R_k(x) = \frac{\sum_{i \in \mathcal{R}:\, g_i \neq g_x} \mathbb{I}[q_k(x_i) \leq q_k(x)]}{1 + \sum_{i \in \mathcal{R}} \mathbb{I}[g_i \neq g_x]}.$$

The density summary used in the main-text fusion equation is

$$D(x) = \tfrac{1}{4}\left(R_{\mathrm{scale}} + R_{\mathrm{shape}} + R_{\mathrm{spectral}} + R_{\mathrm{learned}}\right).$$

Independent baseline suite. The external baselines use the same scored windows and held-out groups as TC-REJECT and are refitted within each fold using only its labeled fit windows. Feature-space methods receive the concatenated standardized scale, shape, and spectral features after fold-specific PCA with at most 200 components; VAR(1) instead uses the raw windows. Fixed settings are

$k = 10$ for kNN and inverse-distance-weighted kNN voting; 200 trees and at most 256 samples for Isolation Forest; 95% explained variance for PCA residual; an RBF One-Class SVM with $\nu = 0.1$, scaled bandwidth, and at most 5,000 fit windows; a pooled ridge VAR(1) with ridge multiplier 0.01; and a 200-tree random forest with minimum leaf size 3 and at most 20,000 fit windows. The first five methods are class agnostic, whereas random-forest uncertainty and kNN voting use fit-window labels. Randomized methods use seed 0, scores are averaged across folds that score a window, and no baseline hyperparameter is tuned by dataset or target.

G AJILE12: PER-PARTICIPANT COHORT TABLE

Structural novelty and the unresolved blend-type target are constructed consistently across all ten participants. The finer activity vocabulary varies by recording day, so class-specific targets without a common definition are not averaged across participants. Table G.1 reports target-window counts and block-bootstrap 95% confidence intervals for both fusion rules. The pooled values are mean $\pm$ standard deviation across participants and are not confidence intervals.

Table G.1: AJILE12 participant-level results. n_N, n_S, and n_B are the numbers of named-reference, structural-target, and blend-target windows. AUROCs are followed by block-bootstrap 95% confidence intervals.

				E-value sum		Rank average	
Participant	n_N	n_S	n_B	Structural	Blend type	Structural	Blend type
sub-01	5,492	9,224	331	**.827** [.805,.846]	**.615** [.547,.688]	.824 [.801,.843]	.580 [.515,.665]
sub-02	2,240	10,728	1,249	.683 [.644,.719]	**.680** [.624,.734]	**.754** [.715,.788]	.605 [.555,.651]
sub-03	8,142	3,879	1,080	.908 [.891,.922]	**.586** [.543,.625]	**.949** [.938,.959]	.552 [.508,.594]
sub-04	4,038	5,202	382	.566 [.530,.600]	.554 [.470,.625]	**.567** [.529,.607]	**.618** [.530,.700]
sub-05	4,454	6,910	893	.864 [.842,.886]	.566 [.515,.611]	**.889** [.867,.907]	**.582** [.526,.626]
sub-06	638	7,020	2,055	.689 [.599,.782]	.630 [.527,.727]	**.790** [.700,.875]	**.738** [.635,.831]
sub-07	486	5,077	155	.688 [.603,.763]	.494 [.345,.647]	**.832** [.762,.894]	**.569** [.413,.724]
sub-08	286	6,547	414	.582 [.508,.649]	.495 [.392,.594]	**.731** [.653,.804]	**.604** [.498,.701]
sub-10	1,783	10,616	314	.854 [.827,.880]	.623 [.565,.677]	**.884** [.860,.910]	**.675** [.614,.734]
sub-11	1,692	6,696	46	.603 [.561,.646]	**.698** [.565,.832]	**.661** [.621,.701]	.642 [.588,.693]
Mean $\pm$ SD	–	–	–	.726 $\pm$ .127	.594 $\pm$ .069	**.788** $\pm$ .114	**.617** $\pm$ .056

The participant-level results show that the average advantage of rank fusion is not universal: its effect depends on both participant and novelty target. The main conclusions are therefore based on cohort summaries and the paired distributions in Figure 3, not on any single recording.

H FRESH AJILE12 CONFIRMATORY EVALUATION

The fusion rules were prespecified before evaluation on 12 previously unused sessions. Structural novelty is available in 11 sessions and blend-type novelty in all 12. The prespecified metric averages the available target AUROCs within each session. Rank-average fusion exceeds e-value-sum fusion by 0.086 AUROC on this metric, is higher in 11 of 12 sessions, and has a subject-bootstrap 95% confidence interval of $[0.049, 0.123]$ for the paired difference.

Table H.1: Mean AUROC on the previously unused AJILE12 sessions.

Method	Structural	Blend type
TC-REJECT, rank average	**0.813**	**0.658**
PCA residual	0.798	0.538
Isolation Forest	0.776	0.546
kNN distance	0.774	0.547
OpenMax-style (spectral)	0.760	0.574
One-Class SVM	0.749	0.539
OpenMax-style	0.733	0.611
TC-REJECT, e-value sum	0.719	0.580
Random-forest uncertainty	0.696	0.623
VAR(1) residual	0.601	0.523
kNN vote	0.579	0.597
Energy	0.491	0.567

I PAMAP2 REPRESENTATION AND EVALUATION DETAILS

Each 10-second window contains the nine raw accelerometer axes from the hand, chest, and ankle sensors. The analysis retains these triaxial signals rather than replacing them by their magnitudes. From the raw axes, SCALE describes amplitude and covariance, SHAPE describes normalized temporal dependence, SPECTRAL uses Welch-averaged log power, and the discriminative and learned channels are constructed as described in Section 3.1. Both fusion rules receive the same channel bank.

Calibration is class conditional: a test window is compared with calibration distances from its nearest named activity rather than with a pooled mixture of activities. Fold-specific class models, learned encoders, and calibration distributions exclude the target participant, and no withheld-activity label is used at scoring time. Run-level hand-crafted-feature standardization, reliability estimation, and adaptive scale weighting are transductive as described in Section 3.1. The baseline methods use the same participant folds and corresponding scale or spectral representations.

Table I.1: PAMAP2 AUROC by held-out activity for the two TC-REJECT fusion rules. Brackets give block-bootstrap 95% confidence intervals.

Held-out activity	E-value-sum fusion	Rank-average fusion
vacuum cleaning	0.798 [0.721, 0.861]	**0.844** [0.779, 0.890]
rope jumping	0.904 [0.841, 0.950]	**0.959** [0.938, 0.989]
Nordic walking	0.972 [0.949, 0.993]	**0.993** [0.986, 0.997]
ascending stairs	**0.889** [0.823, 0.944]	0.888 [0.854, 0.930]
descending stairs	0.962 [0.926, 0.986]	**0.981** [0.963, 0.996]
mean	0.905	**0.933**

Table I.2: PAMAP2 baseline performance on the same participant folds.

Method	Mean AUROC (5 activities)	Description
TC-REJECT, rank-average fusion	**0.933**	common-rank fusion of calibrated channels
OpenMax-style (spectral)	0.929	OpenMax-style score on spectral features
TC-REJECT, e-value-sum fusion	0.905	weighted sum of calibrated e-values
Isolation Forest	0.811	random-partition anomaly score
PCA residual	0.799	subspace-reconstruction error
Energy	0.633	energy score on the shared shallow classifier
VAR(1) residual	0.454	one-step temporal-prediction error

Table I.2 reports the baseline methods displayed in Figure 4C.

J DBS SUPPLEMENTARY ANALYSES

The independently reviewed Extra Movement intervals form four segments in S1 and five in S2. Segment-bootstrap intervals are consequently much wider than a window-level analysis would suggest. Rank-average fusion obtains AUROC 0.774 (95% CI 0.257–0.947) on S1 and 0.689 (0.549–0.877) on S2. The corresponding e-value-sum values are 0.778 (0.346–0.930) and 0.749 (0.660–0.895). Figure 5C reports the baseline results separately by participant because the available segments do not support a stable pooled ranking.

E-value-sum relabeling provides the highest precision in both recordings, but its relabeling recall of 0.390 and 0.250 prevents a claim of complete relabeling. Balanced accuracy for the clinical, e-value-sum, rank-average, and two-score-rank-control training labels is 0.666, 0.754, 0.721, and 0.734 in S1, and 0.319, 0.360, 0.367, and 0.385 in S2. The two-score rank control averages spectral Mahalanobis and maximum-softmax ranks. The corresponding downstream-decoder Extra-Movement recalls are 0.366, 0.463, 0.561, and 0.512 in S1 and 0.500, 0.636, 0.705, and 0.727 in S2. These post-hoc comparisons are descriptive because only two recordings are available and Rest specificity remains low in S2 under every condition.

Table J.1: DBS relabeling rules evaluated only within clinically annotated Rest. Precision measures how often a reassigned window is independently reviewed as Extra Movement; FRR is the fraction of genuine Rest windows incorrectly reassigned.

Recording	Rule	Reassigned	Precision	Recall	FRR
S1	Nearest Movement class	37	0.027	0.024	0.303
S1	LDA Movement posterior	77	0.234	**0.439**	0.496
S1	TC-REJECT, e-value sum	17	**0.941**	0.390	**0.008**
S1	TC-REJECT, rank average	21	0.714	0.366	0.050
S1	Two-score rank control	20	0.500	0.244	0.084
S2	Nearest Movement class	24	0.125	0.068	0.228
S2	LDA Movement posterior	99	0.222	**0.500**	0.837
S2	TC-REJECT, e-value sum	17	**0.647**	0.250	**0.065**
S2	TC-REJECT, rank average	19	0.579	0.250	0.087
S2	Two-score rank control	17	0.412	0.159	0.109

K Additional downstream classification results

K.1 AJILE12 participant-level classification

Table K.1 reports the seven participants with sufficient support for the named-activity classifier. All conditions use identical held-out blocks. The random-removal condition matches the number of windows excluded by label refinement; the two score-filtered conditions use the e-value-sum and rank-average novelty scores.

Table K.1: AJILE12 named-activity balanced accuracy by participant.

Participant	All windows	Matched random	Label refined	E-value filtered	Rank filtered
sub-01	0.658	0.665	**0.679**	0.657	0.654
sub-02	0.809	0.821	0.816	**0.828**	0.821
sub-03	**0.760**	0.757	**0.760**	0.758	0.759
sub-04	0.505	0.505	0.501	**0.507**	**0.507**
sub-05	0.749	0.762	**0.770**	0.761	0.757
sub-06	0.744	0.904	**0.940**	0.754	0.736
sub-07	**0.608**	0.597	0.597	**0.608**	0.545
Mean	0.690	0.716	**0.723**	0.696	0.683

Among the seven participants with sufficient support for the downstream classifier, label refinement improves on the full training set for four participants, with one tie, and improves on matched random removal for four participants, also with one tie. The large gain for sub-06 contributes substantially to the cohort mean, so the average benefit should not be interpreted as uniform.

K.2 PAMAP2 training-budget analysis

Table K.2 reports the finalized split-A training-budget analysis. It reuses the prespecified out-of-fold TC-REJECT scores, applies both fusion rules at $\alpha = 0.10$, averages five random seeds, and evaluates the eight eligible held-out participants defined in Appendix D with random forests and XGBoost. Each entry compares score-informed removal with random removal of the same number of windows.

Table K.2: PAMAP2 paired balanced-accuracy differences between score-informed and size-matched random removal. Values are means over five seeds and eight held-out participants, followed by participant-bootstrap 95% confidence intervals. Positive values favor score-informed removal; bold indicates an interval that excludes zero.

	Random forest		XGBoost	
Named-window budget	E-value sum	Rank average	E-value sum	Rank average
Full	.001 [-.001,.004]	**-.005 [-.009,-.001]**	.005 [-.001,.010]	-.010 [-.021,.0004]
1500	**.003 [.001,.005]**	-.004 [-.009,.002]	**.007 [.0001,.014]**	.004 [-.007,.016]
750	.003 [-.0001,.006]	-.004 [-.011,.003]	**.006 [.002,.011]**	.006 [-.005,.019]
300	.004 [-.0004,.008]	-.002 [-.016,.008]	.009 [-.002,.023]	-.004 [-.030,.024]
150	.005 [-.001,.011]	.019 [-.0002,.033]	-.004 [-.014,.005]	.006 [-.012,.025]
100	**.011 [.004,.020]**	**.026 [.005,.047]**	**.017 [.004,.030]**	.011 [-.017,.037]
75	.006 [-.002,.015]	**.034 [.006,.063]**	-.006 [-.027,.015]	**.052 [.018,.090]**
50	-.008 [-.022,.004]	-.001 [-.020,.020]	.001 [-.007,.009]	.009 [-.016,.033]

Rank-average removal is beneficial relative to matched random removal at 75 windows for both classifiers. At 100 windows, the random-forest difference remains positive with an interval excluding zero, whereas the XGBoost interval includes zero. No rank-average advantage is evident at budgets of at least 150 windows or at 50 windows. E-value-sum differences are smaller and vary across budgets. A label-fit rule was also evaluated, but its apparent random-forest advantage at 100 windows did not replicate with XGBoost, so it is not included in the benefit claim. These results are descriptive because they use one withheld-activity split, one pinned score run, five seeds, and eight evaluable participants.