DISENTANGLING OOD DETECTION AND MODEL UNCERTAINTY IN HIGH-RISK EEG APPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning models for electroencephalography (EEG) analysis show great promise across a wide range of applications, but their deployment in high-risk domains is hindered by their vulnerability to distribution shifts. Encountering out-of-distribution (OOD) data can lead to catastrophic, overconfident predictive failures. While OOD detection methods can mitigate these risks, they remain heavily under-explored for EEG. Moreover, evaluations in the broader literature typically assess OOD detection performance in isolation, ignoring their practical impact on downstream applications. To bridge this gap, we introduce a benchmark for EEG OOD detection, evaluate a broad range of methods, and investigate their value in two downstream clinical prediction tasks. Our results disentangle OOD detection and model uncertainty estimation capabilities, which are frequently conflated in the literature, provide actionable insights about the current state of the art for EEG OOD detection and model uncertainty estimation, and demonstrate how complementary methods for both aspects can be combined to form a robust safety net for the deployment of machine learning models in real-world applications.

1 INTRODUCTION

Machine learning (ML) provides a promising avenue for supporting and enhancing the analysis of electroencephalography (EEG). However, especially in high-risk contexts like clinical applications, the use of machine learning-based EEG biomarkers is hindered by their vulnerability to distribution shifts. Machine learning models can fail catastrophically (and with high confidence) when they are presented with data that differs substantially from the data they were trained on, i.e. out-of-distribution (OOD) data (Nguyen et al., 2015). Particularly for EEG, there are ample ways in which such shifts can be introduced, including different recording hardware, (accidental) mismatches in preprocessing (e.g. filter settings or re-referencing), or differences in the patient population. While specific quality controls for some of these can be designed, part of the problem is that there can be unanticipated shifts. Furthermore, while for example individual differences between patients can be the cause of distribution shifts (Bomatter & Gouk, 2025), only deploying a model for patients that the model has seen during training is often not an option. In such cases, it may not be clear a priori whether the data from a new patient is similar enough to the training distribution.

OOD detection can mitigate these risks by enabling a system to abstain from making a prediction or defer to a human expert when a case is likely out of distribution. Current OOD detection methods can broadly be categorised into two paradigms: discriminative and generative. Approaches such as Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017), Energy scores (Liu et al., 2020), Out-of-Distribution Detector for Neural Networks (ODIN) (Liang et al., 2018), and Activation Shaping (ASH) (Djurisic et al., 2023) directly operate on the activations or outputs of a (discriminative) classification model. Conversely, methods based on a generative model—including likelihood estimates, Typicality (Nalisnick et al., 2019a), Density of States Estimation (DoSE) (Morningstar et al., 2021), and Signal in the Noise (SITN) (Bomatter et al., 2026)—attempt to explicitly model the underlying distribution of the training data. While these methods have seen extensive development and benchmarking in domains like computer vision, their application to EEG data remains heavily under-explored. Furthermore, the standard evaluation paradigm in the broader OOD detection literature suffers from a significant disconnect from clinical utility. The vast majority of existing work evaluates OOD detection methods strictly in isolation, assessing only how well a method can separate an in-distribution (ID) dataset from a distinctly different OOD dataset (e.g.

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.

separating CIFAR-10 images from SVHN images). Crucially, such evaluations fail to reflect the downstream impact of OOD detection: when deploying a system for clinical diagnosis, the ultimate goal is not merely to detect OOD samples, but to understand how flagging such samples improves the reliability of the diagnosis.

To bridge this gap, this work presents a systematic evaluation of OOD detection for EEG, explicitly linking detection capabilities to downstream performance in clinical prediction tasks. We design a suite of practically relevant perturbations with different levels of severity and use them to benchmark a broad set of OOD detection methods spanning the aforementioned approaches based on both discriminative and generative models. In line with a recent theoretical critique and position paper (Li et al., 2025b;a), our results disentangle two fundamentally distinct concepts that are frequently conflated in the literature: true OOD detection vs model uncertainty. Across perturbations, methods based on discriminative models were largely incapable of OOD detection, while methods based on generative models achieved considerably better results, with increasing performance for higher perturbation severities. However, in downstream clinical evaluations—where model errors stem from both OOD data and difficult in-distribution cases—we found that methods from both paradigms effectively predicted model performance. While generative models recognise unfamiliar data, discriminative methods can serve as a valuable proxy for model uncertainty. Ultimately, we show that these complementary signals can be combined to form a more robust safety net for the deployment of EEG-based machine learning models in clinical and other high-risk applications.

In summary, our main contributions are as follows:

- We present an OOD detection evaluation framework for EEG that closes several gaps in the field. The framework comprises a suite of perturbations for the controlled creation of OOD EEG data, as well as investigations of downstream impacts in two clinical prediction tasks.
- We benchmark a broad range of OOD detection methods, including both discriminative and generative approaches, the latter of which have not previously been probed in an EEG context, but achieved substantially better performance.
- We present strong empirical evidence that clearly disentangles the often conflated concepts of OOD detection and model uncertainty, a finding and experimental setup that are of broader interest to the OOD detection community beyond EEG.
- Our results provide actionable guidance on the current state of the art for EEG OOD detection and show how it can be combined with uncertainty estimation for the safer deployment of machine learning models in high-risk EEG applications.

2 Related Work

OOD Detection for EEG-based Machine Learning The EEG-specific literature on OOD detection is limited. In a recent preprint, Mulder et al. (2026) benchmarked several discriminative OOD detection methods in a motor imagery setting and found them to be largely ineffective. Liu et al. (2026) proposed a temporal OOD detection method for EEG-based brain-computer interfaces (BCIs) and compared it against a range of standard discriminative OOD detection methods. Our study investigates a more general setting than the latter and is focused on a clinical context rather than BCIs. Furthermore, while our results largely concur with their findings on discriminative methods, our experiments also include methods based on generative models, an important addition given that they displayed much better OOD detection capabilities. Concurrent work by Tveter et al. (2026) also introduced a range of perturbations to create EEG data with controlled distribution shifts. However, their work is focused on the comparison of different ensemble learning methods and their robustness to distribution shifts rather than OOD detection.

OOD Detection Benchmarks in Other Domains Outside of EEG, the OOD detection literature is particularly advanced in computer vision. Notable works that have focused on the evaluation of different OOD detection methods are the OpenOOD benchmarks (Yang et al., 2022; Zhang et al., 2023), although many methods papers also include comprehensive comparisons against other methods (Hendrycks & Gimpel, 2017; Liu et al., 2020; Liang et al., 2018; Djurisic et al., 2023; Morningstar et al., 2021; Bomatter et al., 2026). Less mature, but more closely related to EEG, is the field of OOD detection for (more general) time series data. Within this field, the OOD detection benchmark for

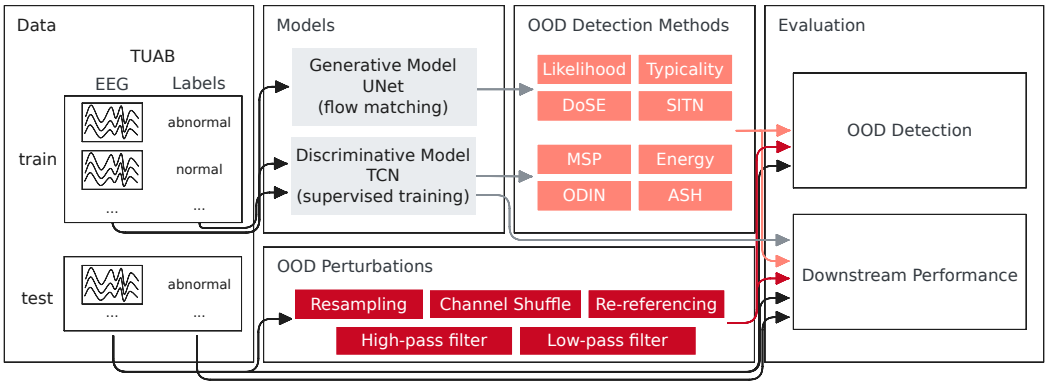

Figure 1: **OOD detection evaluation concept.** An EEG dataset is partitioned into training and test splits. The training data is used to train two models: a generative model trained via unsupervised learning (here a UNet trained via flow-matching), and a discriminative model trained through supervised learning (here a TCN model). The test split serves as in-distribution test data, for which corresponding out-of-distribution data is generated through controlled perturbations. Finally, we evaluate a range of OOD detection methods based on either the generative or discriminative model. In addition to OOD detection performance, we also evaluate its impact on downstream performance, quantifying how the prediction accuracy of the discriminative model varies across samples with different OOD scores of the different methods.

time series data by Gungor et al. (2025) is most relevant to our work. However, like EEG-specific prior work, their benchmarks are limited to discriminative approaches, for which they reported relatively poor performance overall. Furthermore, a common thread across all these works is their primary reliance on semantic shifts to define OOD data, such as reserving a subset of classes as OOD or evaluating models across disjoint datasets. Guille-Escuret et al. (2024) recently criticised these approaches, arguing that an exclusive focus on semantic shifts neglects relevant (covariate) distribution shifts encountered by real-world systems. Motivated by a practical perspective, we see the value of OOD detection in protecting against model failure, irrespective of the nature of the shift. Our work is therefore closely aligned with this view. Our work is also different in that we go beyond standard OOD detection evaluations to study impacts on downstream clinical applications and disentangle the concepts of true OOD detection and model uncertainty.

3 Methods

Figure 1 provides an overview of the components that enable our EEG OOD detection evaluations. Starting from an EEG dataset with corresponding annotations for a (clinical) prediction task, samples are partitioned into training and test splits. The training data is used to train two models: a generative model trained using only the EEG data without labels (i.e. unsupervised), and a discriminative model trained through supervised learning. The test set then serves as in-distribution test data, for which we generate corresponding out-of-distribution data through controlled perturbations. Finally, we evaluate a range of OOD detection methods based on either the generative or discriminative model in two ways. First, we assess their OOD detection performance, i.e. how well their assigned OOD scores discriminate between the unperturbed ID test samples and their corresponding perturbed OOD versions. Second, we evaluate the impact of OOD detection on downstream performance in (clinical) prediction tasks, quantifying how the prediction accuracy of the discriminative model varies across samples with different OOD scores of the different methods.

3.1 OOD Perturbation Suite for EEG

We introduce a suite of controlled perturbations to create OOD samples from in-distribution test data. These transformations were designed to simulate plausible distribution shifts that may be encountered in real-world EEG data, though they are fundamentally intended to serve as a structured,

reproducible, and well-controlled testbed rather than an exhaustive modelling of reality. Furthermore, the perturbations were designed to be parametrisable, which (with the exception of the re-referencing perturbation) enables fine-grained control of the severity of the distribution shift. Finally, by applying these transformations directly to the ID test set, the resulting OOD data is perfectly paired with the unperturbed baseline, providing the methodological capacity to isolate the exact effect of a shift on model performance. See Figure 2 for an illustration of all the perturbations.

Sampling Rate The signal is resampled to a higher sampling rate (using bandlimited sinc interpolation via the torchaudio library) and then cropped to maintain the same number of samples. From the perspective of a model that expects data to be at a lower sampling rate, perturbed samples will appear smoothed. Sampling rate mismatches are surprisingly common in practice; for instance, a common issue in the evaluation of EEG foundation models is that they are applied to data with a different sampling rate than the model was pre-trained on. We control the severity of this perturbation by varying the amount by which the sampling rate is increased compared to the 100 Hz that the model was trained on, ranging from 125 Hz through 150 Hz to 200 Hz.

Channel Shuffle The ordering of a subset of EEG channels is permuted. From the perspective of a model that was trained with a fixed channel order, the perturbation will lead to strange topographies with, for example, eye blinks appearing in presumed occipital channels. In practice, perturbations like this could be introduced by accident with some recording devices where electrodes are plugged into the amplifier individually. We control the severity of this perturbation by varying the size of the subset of randomly selected channels which are permuted. We increase the subset from 10%, i.e. swapping just two channels in the 19-channel layout, to 50% and then 100%.

Reference Scheme The EEG channels are re-referenced to a different scheme. Starting from the average-referenced data used to train the model, we re-reference to three different schemes: a monopolar vertex reference with Cz as the reference channel, a linked-temporal reference with the average of T7 and T8 as the reference, and a longitudinal double-banana bipolar montage with an additional transverse channel (T7-T8) to maintain the same channel count. Unlike the different levels of other perturbations, these options do not have a clear ordering in terms of perturbation severity.

High- and Low-pass Filters The signal is filtered with an additional zero-phase Butterworth high- or low-pass filter. For the high-pass filter perturbation, we start from the unperturbed data (high-pass filtered at 1 Hz) and vary the severity of the perturbation by increasing the cutoff frequency to 2 Hz, 4 Hz, and 8 Hz, progressively removing more low-frequency activity. For the low-pass filter perturbation, the severity is varied by lowering the cutoff frequency from the 50 Hz in the unperturbed data to 45 Hz, 30 Hz, and 15 Hz, progressively removing more high-frequency activity.

3.2 OOD Detection Methods

Discriminative methods The discriminative methods included in the benchmarks were selected to broadly cover different established strategies in the literature. For classical output-based scoring, we include the maximum softmax probability (MSP) (Hendrycks & Gimpel, 2017) as well as energy scores computed from unnormalised logits (Liu et al., 2020) and the more recent LogitGap (Liang et al., 2025). Out-of-Distribution Detector for Neural Networks (ODIN) (Liang et al., 2018) is an intervention-based technique that applies temperature scaling to the model's logits prior to the softmax computation and adds gradient-based perturbations to the input data, which is expected to have a stronger impact on ID than OOD data and improve their separation. ASH (Djurisic et al., 2023) is an activation shaping method that intervenes in the forward pass of a model by pruning and transforming activations before they reach the final classification head. Finally, the Mahalanobis distance (Lee et al., 2018) represents a feature-based approach.

For ODIN, we adopt the hyperparameters reported in the original publication for the main results ($T = 1000$, $\epsilon = 0.0014$). For ASH, we specifically use the ASH-S version with a pruning percentage of 65%. A hyperparameter sensitivity analysis was performed (see Appendix H). As the underlying discriminative model for all methods, we use the TCN architecture (Bai et al., 2018), which has been adapted and is widely used for EEG data (Gemein et al., 2020) (see Appendix G for hyperparameters and training details).

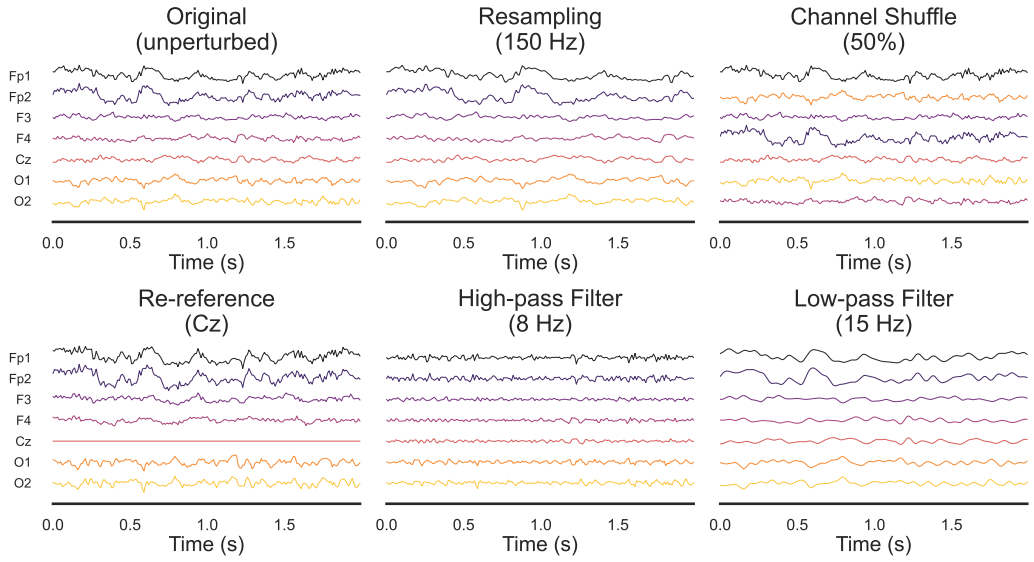

Figure 2: **Illustration of EEG OOD perturbations.** An unperturbed (ID) EEG segment (top-left) is shown along with corresponding OOD samples obtained through the different perturbations.

Generative methods Among generative approaches, we first include the standard log-likelihood. Furthermore, to address failure modes of likelihoods in high dimensions, Typicality (Nalisnick et al., 2019a) measures the absolute difference between a sample's negative log-likelihood and the expected ID entropy. Density of States Estimation (DoSE) (Morningstar et al., 2021) instead evaluates the empirical density of multiple statistics. We use the DoSE_{KDE} version for flow-based models, which is based on three statistics: the log-likelihood, the log-probability of the latent representation, and the integral of the vector field's divergence. Finally, Signal in the Noise (SITN) (Bomatter et al., 2026) leverages the diffeomorphic properties of continuous normalising flows, flagging OOD samples if their backward-integrated noise representations deviate from the training prior.

In-distribution validation data is used to fit the entropy estimate for Typicality, the KDE for DoSE, and the max-quantile transform for SITN. As the underlying generative model for all methods, we use the UNet architecture following Ho et al. (2020), as implemented by Dhariwal & Nichol (2021), and train it using the flow matching objective by Lipman et al. (2023) (see Appendix G for details). We use Euler's method (with 100 timesteps), implemented in the library by Lipman et al. (2024), as the ODE solver to obtain the likelihoods and noise samples required by the different methods.

3.3 Data

Experiments were conducted on two large datasets (TUAB, CAUEEG) with annotations for clinical prediction tasks. With more than 1,000 participants each, these datasets are among the largest EEG datasets available for research purposes, making them particularly suitable for robust training and benchmarking of machine learning methods in our study.

TUAB The Temple University Hospital Abnormal (TUAB) EEG Corpus (López et al., 2015; Lopez de Diego, 2017) comprises a demographically balanced and curated subset of more than 2,300 participants from the TUH EEG Corpus (Obeid & Picone, 2016). The data was collected in a hospital environment and recordings were annotated by board-certified neurologists as normal (1385 participants) or abnormal (998 participants) according to standardised criteria (Rubin, 2021). We used version 3.0.1 of the dataset and followed the data curation and splitting procedure in (Bomatter & Gouk, 2025), which respects the official train-test split and additionally defines a validation split with 10% of the training data. All samples of a given participant were assigned to the same split. Data access requests should be directed to the dataset authors.

CAUEEG The Chung-Ang University Hospital EEG (CAUEEG) dataset (Kim et al., 2023) includes diagnostic labels for the categories normal (459 participants), mild cognitive impairment (MCI; 417 participants), or dementia (311 participants). We used the official split without overlapping participants. Data access for research purposes can be requested from the dataset authors.

Preprocessing The data preprocessing comprised the following steps: selection of a fixed channel subset (the 19 channels in the 10–20 system), band-pass filtering to 1-50 Hz, resampling to 100 Hz, and re-referencing to average reference. Finally, non-overlapping epochs with 2s duration were used.

4 RESULTS

4.1 OOD DETECTION PERFORMANCE

Table 1 shows the Area Under the Receiver Operating Characteristic (AUROC) curve for the OOD detection performance across the different perturbations and severity levels for all methods on the TUAB dataset. Models for the discriminative and generative methods were trained on the train split of the (unperturbed) dataset and OOD detection methods were then evaluated using the unperturbed test split of the dataset as ID samples and the perturbed test data as OOD samples.

Table 1: **OOD detection performance on TUAB.** Performance is reported in terms of AUROC and bold values highlight the best result for each perturbation and severity level. Discriminative-model-based methods (MSP, Energy, ODIN, ASH) perform largely at chance level, whereas generative-model-based methods achieve considerably better OOD detection with increasing performance for higher severity levels. SITN consistently achieves the best performance.

Perturbation	Severity	MSP	Energy	ODIN	ASH	LL	Typ.	DoSE	SITN
Resampling	125 Hz	0.517	0.500	0.514	0.533	0.369	0.559	0.556	**0.947**
	150 Hz	0.457	0.479	0.453	0.534	0.275	0.646	0.647	**0.981**
	200 Hz	0.383	0.451	0.377	0.522	0.143	0.816	0.826	**0.989**
Channel Shuffle	10%	0.549	0.494	0.549	0.513	0.607	0.521	0.543	**0.754**
	50%	0.553	0.431	0.555	0.494	0.793	0.682	0.711	**0.924**
	100%	0.526	0.391	0.529	0.460	0.823	0.726	0.748	**0.942**
Re-referencing	Cz	0.517	0.435	0.520	0.597	0.950	0.937	0.987	**0.998**
	T7, T8	0.587	0.513	0.591	0.568	0.927	0.896	0.973	**0.997**
	bipolar	0.752	0.621	0.757	0.719	0.940	0.919	0.962	**0.996**
High-pass Filter	2 Hz	0.514	0.558	0.511	0.506	0.475	0.504	0.503	**0.604**
	4 Hz	0.531	0.603	0.526	0.550	0.445	0.510	0.510	**0.749**
	8 Hz	0.446	0.684	0.439	0.637	0.367	0.546	0.544	**0.872**
Low-pass Filter	45 Hz	0.499	0.502	0.499	0.503	0.461	0.507	0.505	**0.822**
	30 Hz	0.500	0.511	0.498	0.515	0.246	0.649	0.644	**0.986**
	15 Hz	0.511	0.536	0.506	0.560	0.016	0.975	0.974	**0.990**

OOD detection methods operating on discriminative models that were trained in a supervised way on the normality prediction task (MSP, Energy, ODIN, ASH) were largely ineffective for differentiating between ID and OOD data. These methods performed close to chance level (AUROC of 0.5) for most conditions and showed inconsistent trends with perturbation severity. Energy scores, for instance, showed increasingly better OOD detection performance with higher severities of the high-pass filter perturbation, but decreasing performance for increasing severities of the resampling perturbation.

Generative models (in particular Typicality, DoSE, and SITN), on the other hand, achieved much better OOD detection overall with consistent severity-dependent performance improvements. SITN clearly emerged as the most effective OOD detection method with strong performance even at the lowest severity level and for perturbations where other methods performed close to chance level (see for example the channel shuffle perturbation at the lowest severity level where only two channels are swapped). Raw log-likelihoods worked well for certain perturbations (Channel Shuffle, Re-referencing), but not others, exhibiting a behaviour previously observed in the computer vision literature, which we discuss in more detail in Appendix F.

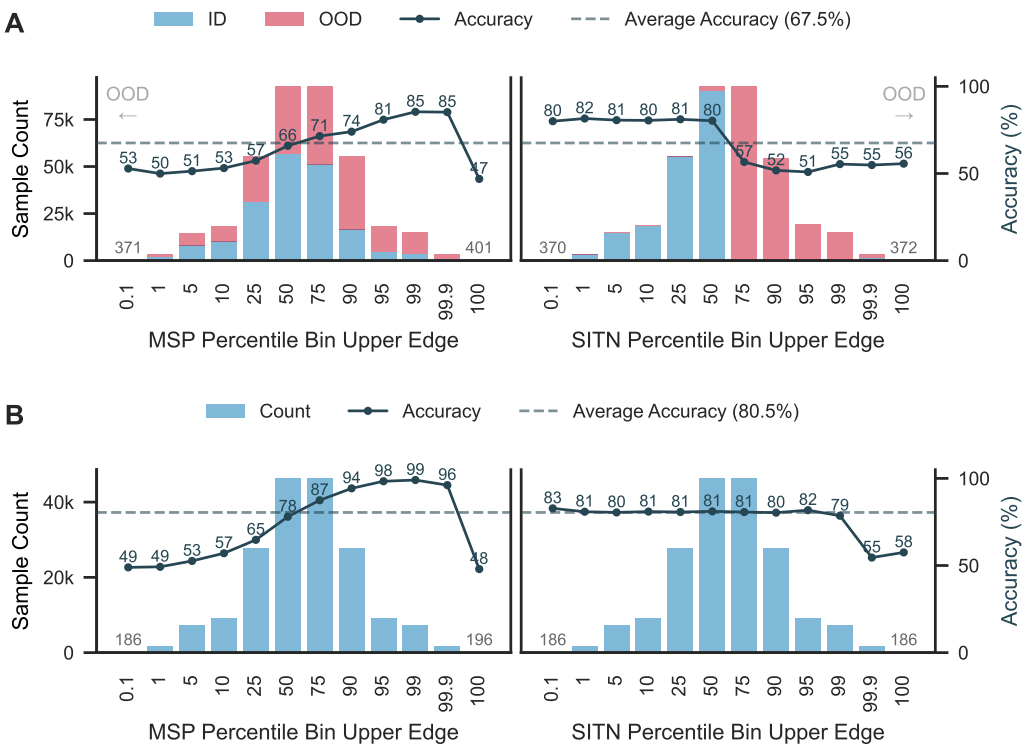

Figure 3: **Model performance for normality prediction on the TUAB dataset.** Accuracy across percentile bins, where the x-axis ticks indicate the upper edge of the corresponding bin (i.e. the first bin contains samples from the 0-0.1 percentile, the second one 0.1-1, etc.). Histograms indicate the number of ID and OOD samples in each bin. **A** Evaluation on ID and corresponding OOD samples obtained with the resampling perturbation (severity 200 Hz). **B** Evaluation on ID-only data.

Evaluations on CAUEEG corroborate these results, except that DoSE outperformed SITN in some settings (see Appendix B). Additional results for Mahalanobis distance and LogitGap can be found in Appendices C and D. Appendix A shows an extended version of Table 1 with confidence intervals.

4.2 Downstream Performance Impact

Panel A in Figure 3 shows the impact of OOD detection on the clinical downstream EEG normality prediction task on the TUAB dataset (see Section 3.3). The figure shows one method from each paradigm (MSP as a discriminative method and SITN as a generative method) using the resampling perturbation (severity 200 Hz) for the OOD data. For each method, the pooled in- and out-of-distribution samples were sorted according to the assigned OOD scores. Then we report the (downstream) classification accuracy of the discriminative model across different percentile bins of the OOD score (e.g. the performance across the samples with the lowest 0.1% MSP scores, the 0.1-1% lowest MSP scores, etc.). The arrows in the top-left/right corners of the plots indicate the OOD direction for each method: for MSP, OOD samples are expected to have lower scores, whereas SITN should assign higher scores to OOD samples. Finally, the histograms show the sample counts in each percentile bin with the different colours indicating the proportion of ID and OOD samples.

For MSP, the proportion of ID and OOD samples is relatively similar across bins, consistent with the poor OOD detection results reported in Section 4.1. MSP does not separate ID and OOD samples well. Despite this fact, MSP values are still predictive of model performance, except for a failure mode at very high confidence values. SITN, on the other hand, separates ID and OOD samples extremely well. It can be seen that the model performance is substantially lower on OOD data, demonstrating the value of OOD detection as a safety net to prevent misclassifications.

In Panel B, we present the same visualisations on ID data alone. It can be seen that MSP scores are much more predictive of model performance than SITN scores within ID samples. Generative approaches like SITN, which do not have access to the discriminative model used for the classification task, provide little indication of model performance within the majority of ID samples, but they do not have the aforementioned failure mode of discriminative methods at very high confidence levels. A performance drop can be observed for ID samples with the most extreme (top 1%) SITN scores, which likely represent rare ID cases to which the model had very little exposure during training.

Overall, these results clearly disentangle two aspects that influence downstream performance. Generative methods like SITN excel at true OOD detection, identifying samples outside the training distribution, for which model predictions should not be trusted because the model was not exposed to comparable data during training. Discriminative methods like MSP, on the other hand, are not suitable for OOD detection, but reflect model uncertainty. Even with ample exposure to similar samples during training, certain cases can be ambiguous and difficult for the model to classify. Model uncertainty can capture such cases and provide valuable information about how well the model is likely to perform across different in-distribution samples. Full results for all perturbations and OOD detection methods as well as evaluations on the dementia diagnosis task on CAUEEG can be found in Appendix L. While the effect on downstream performance was smaller for certain conditions (e.g. Low-pass filter on CAUEEG; Figure 22), the general patterns are highly consistent.

4.3 Combination of OOD Detection and Model Uncertainty

Based on the observations in Section 4.2, we evaluated a system in which samples are assessed by both a (generative) OOD detector and a (discriminative) model uncertainty estimator. The classifier then only provides predictions for samples that pass both tests and abstains from prediction otherwise. We used SITN for OOD detection, following the calibration procedure detailed in the original publication (Bomatter et al., 2026), which allows one to pick a threshold with access only to an ID validation set based on an acceptable false positive rate (FPR). A target FPR of 1% was chosen for this example. To control model uncertainty, we first used isotonic regression (fitted on the ID validation set) to map raw MSP values to calibrated probabilities. We then rejected samples as too uncertain if the calibrated probability of the prediction being correct was below 60%. The choice of this threshold will in practice depend on the costs of misclassifications vs abstaining from prediction.

Results on TUAB are shown in Table 2. Focusing on the baseline performance without any rejections (coverage 100%) first, it can be seen that the accuracy dropped from 80.5% on ID data to 67.5% when corresponding OOD samples (resampling perturbation at 200 Hz) were included in the evaluation. When SITN was used to reject OOD samples, performance remained stable across both conditions. Note also that, as expected, the coverage on ID data is close to the desired 1% FPR. Using MSP to reject high-uncertainty samples improved performance on ID data (at the cost of a lower coverage). When OOD samples were included in the evaluation, MSP did not protect well against the drop in performance, with only marginally better accuracy than the baseline despite rejecting more than 10% of the samples. Combining the two methods resulted in the best accuracy across both conditions.

Table 2: **Selective classification performance for OOD detection and model uncertainty estimation.** OOD detection (using SITN), model uncertainty (using MSP), and their combination (SITN+MSP) are compared on ID-only data and ID+OOD data. OOD data was generated with the resampling perturbation at 200 Hz. MSP provides an accuracy improvement on in-distribution data, but only SITN provides effective protection against OOD data. Combining both methods enables the highest accuracy in both the ID-only and ID+OOD conditions.

	ID Only		**ID and OOD**	
Rejection Method	**Coverage**	**Accuracy**	**Coverage**	**Accuracy**
Baseline (No Rejection)	100.0%	80.5%	100.0%	67.5%
SITN (OOD Detection)	98.5%	80.8%	49.5%	80.7%
MSP (Model Uncertainty)	88.1%	83.8%	89.2%	69.3%
Combined (SITN + MSP)	86.9%	84.1%	43.7%	84.1%

5 Discussion and Conclusions

In this study, we introduced an OOD detection benchmark for EEG-based machine learning and systematically evaluated a broad spectrum of both discriminative and generative methods. Our results clearly demonstrate that standard discriminative methods are incapable of reliably detecting OOD EEG data under realistic covariate shifts, a finding that aligns with limited prior work in the EEG BCI literature (Mulder et al., 2026; Liu et al., 2026). However, generative approaches—which to the best of our knowledge have not previously been benchmarked for EEG or general time-series OOD detection—yield substantially better performance.

Follow-up evaluations to assess the utility of OOD detection in two clinical downstream prediction tasks demonstrated two things. First, performance dropped substantially on out-of-distribution samples, emphasising the value of OOD detection to prevent mispredictions in high-risk applications. Second, both discriminative and generative methods were predictive of downstream performance, but for different reasons. Only generative methods protect against model failure by flagging OOD samples. Discriminative methods prevent mispredictions by highlighting samples with high model uncertainty, even if they are in-distribution. These results provide strong support for the position recently put forward by Li et al. (2025b), which argues that discriminative methods "answer the wrong question for OOD detection". We furthermore demonstrated how the complementary capabilities of discriminative and generative methods can be combined to protect against both failure modes.

The present study has some limitations that highlight valuable directions for future work. While the creation of OOD data through the perturbations presented in this work has several advantages (unambiguous ground truth, perfect pairing with ID samples, large sample size, control of severity levels), the field would benefit from additional benchmarks and curated datasets to study a wider variety of distribution shifts. Designing such OOD detection setups is more challenging than in computer vision. While images from different datasets (e.g. CIFAR-10 vs SVHN) or different object categories (e.g. cats vs cars) are clearly distinct, comparing EEG segments across different datasets is more nuanced. Besides clear differences in recording hardware or preprocessing, aspects of which are reflected in our perturbation experiments, the degree to which differences in patient populations manifest in EEG distribution shifts is hard to quantify. Even for comparatively well-studied variables like sex or age with known population-level differences (Aurlien et al., 2004; Chiang et al., 2011; Cellier et al., 2021), much of the variance is driven by individual differences. Moreover, cross-dataset label definitions are often not mutually exclusive. EEG segments in the TUAB dataset could in principle also be assigned a label for the dementia diagnosis task in CAUEEG, such that CAUEEG segments need not necessarily be OOD for a TUAB-trained model (see exploratory analysis in Appendix J). Where possible, we recommend grounding OOD detection through simultaneous assessments of performance in a relevant downstream task, as in our experiments. This ensures that progress in OOD detection will translate to superior safeguards in practical applications.

Our results also raise questions about benchmarking practices in the broader OOD detection literature. How can the poor performance of discriminative methods in our evaluations be reconciled with their success on widely used computer vision benchmarks like OpenOOD? We hypothesise that this discrepancy stems from fundamental differences in how OOD evaluation is framed. For one, OOD detection benchmarks are dominated by semantic shifts (Guille-Escuret et al., 2024), in which OOD samples correspond to unseen class labels. An influential survey paper by Yang et al. (2024) even frames this exclusive focus as part of the general definition of OOD detection, although it is acknowledged that other shifts may be considered OOD from a generalisation perspective or when considering high-risk applications. Such benchmark designs may favour discriminative methods because the class labels used for training represent the same semantic distinctions that later determine whether a sample is OOD. Model uncertainty, therefore, happens to be aligned with this particular kind of shift. In our setting, on the other hand, distribution shifts are unrelated to the labels of the prediction task. In fact, as in many clinical applications, the label space is exhaustive (e.g. disease present or disease absent) such that there are no unseen classes. In this setup, our results reveal that OOD detection and model uncertainty estimation are fundamentally distinct capabilities, and conflating the two obscures what a given method actually achieves. We encourage future applied work to place practical utility at its centre. In clinical domains, the primary objective of OOD detection is to prevent model failure caused by unanticipated distribution shifts rather than the special case of semantic shifts. Moreover, both true OOD detection and model uncertainty have value. By preventing their entanglement in benchmarks, the field can make targeted progress on both fronts.

AI USE STATEMENT

In this work, generative AI tools were used to assist in the implementation of methods and to support data analysis by assisting in the implementation of visualisation and plotting code. Generative AI tools were not used to generate synthetic data sets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, design experiments, clean and reformat datasets, or interpret results. Formulating mathematical claims, assisting with proofs, and translating text are not applicable to this work. All AI-assisted work was extensively reviewed; specifically, AI-assisted code was manually verified, refined, and tested for correctness by the authors. The authors take full responsibility for the final content of this work, including text, claims, or artefacts produced with the aid of generative AI.

REFERENCES

Hamdi Altaheri, Ghulam Muhammad, and Mansour Alsulaiman. Physics-Informed Attention Temporal Convolutional Network for EEG-Based Motor Imagery Classification. *IEEE Transactions on Industrial Informatics*, 19(2):2249–2258, February 2023. ISSN 1941-0050. doi: 10.1109/TII.2022.3197419. URL https://ieeexplore.ieee.org/document/9852687.

Bruno Aristimunha, Pierre Guetschel, Martin Wimpff, Lukas Gemein, Cedric Rommel, Hubert Banville, Maciej Sliwowski, Daniel Wilson, Simon Brandt, Théo Gnassounou, Joseph Paillard, Bruna Junqueira Lopes, Sara Sedlar, Thomas Moreau, Sylvain Chevallier, Alexandre Gramfort, and Robin Tibor Schirrmeister. Braindecode: toolbox for decoding raw electrophysiological brain data with deep learning models. December 2025. doi: 10.5281/zenodo.17699192. URL https://zenodo.org/records/17699192.

H Aurlien, I. O Gjerde, J. H Aarseth, G Eldøen, B Karlsen, H Skeidsvoll, and N. E Gilhus. EEG background activity described by a large computerized database. *Clinical Neurophysiology*, 115 (3):665–673, 2004. ISSN 1388-2457. doi: 10.1016/j.clinph.2003.10.019.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv preprint abs/1803.01271*, 2018.

Philipp Bomatter and Henry Gouk. Is Limited Participant Diversity Impeding EEG-based Machine Learning? In *Proc. of NeurIPS*, volume 38, pp. 117447–117474, 2025.

Philipp Bomatter, Jack Geary, and Henry Gouk. The Signal in the Noise: OOD Detection Through Goodness-of-Fit Testing in Factorised Latent Spaces. *ArXiv preprint abs/2605.22496*, 2026.

Dillan Cellier, Justin Riddle, Isaac Petersen, and Kai Hwang. The development of theta and alpha neural oscillations from ages 3 to 24 years. *Developmental Cognitive Neuroscience*, 50:100969, 2021. ISSN 1878-9293. doi: 10.1016/j.dcn.2021.100969.

A. K. I. Chiang, C. J. Rennie, P. A. Robinson, S. J. van Albada, and C. C. Kerr. Age trends and sex differences of alpha rhythms including split alpha peaks. *Clinical Neurophysiology*, 122(8): 1505–1517, 2011. ISSN 1388-2457. doi: 10.1016/j.clinph.2011.01.040.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Proc. of NeurIPS*, pp. 8780–8794, 2021.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *Proc. of ICLR*, 2023.

Lukas A. W. Gemein, Robin T. Schirrmeister, Patryk Chrabaszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of EEG pathology. *NeuroImage*, 220:117021, 2020. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2020.117021.

Charles Guille-Escuret, Pierre-André Noël, Ioannis Mitliagkas, David Vázquez, and João Monteiro. Expecting the unexpected: Towards broad out-of-distribution detection. In *Proc. of NeurIPS*, 2024.

Onat Gungor, Amanda Sofie Rios, Nilesh Ahuja, and Tajana Rosing. TS-OOD: Evaluating Time-Series Out-of-Distribution Detection and Prospective Directions for Progress. *ArXiv preprint abs/2502.15901*, 2025.

Jakob Drachmann Havtorn, Jes Frellsen, Søren Hauberg, and Lars Maaløe. Hierarchical vaes know what they don't know. In *Proc. of ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4117–4128, 2021.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proc. of ICLR*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*, 2020.

Hamidreza Kamkari, Brendan Leigh Ross, Jesse C. Cresswell, Anthony L. Caterini, Rahul G. Krishnan, and Gabriel Loaiza-Ganem. A geometric explanation of the likelihood OOD detection paradox. In *Proc. of ICML*, 2024.

Min-jae Kim, Young Chul Youn, and Joonki Paik. Deep learning-based EEG analysis to classify normal, mild cognitive impairment, and dementia: Algorithms and dataset. *NeuroImage*, 272: 120054, 2023. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2023.120054.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Why normalizing flows fail to detect out-of-distribution data. In *Proc. of NeurIPS*, 2020.

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A Compact Convolutional Network for EEG-based Brain-Computer Interfaces. *Journal of Neural Engineering*, 15(5):056013, October 2018. ISSN 1741-2560, 1741-2552. doi: 10.1088/1741-2552/aace8c. URL http://arxiv.org/abs/1611.08024. arXiv:1611.08024 [cs, q-bio, stat].

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proc. of NeurIPS*, pp. 7167–7177, 2018.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Out-of-Distribution Detection Methods Answer the Wrong Questions. *ArXiv preprint abs/2507.01831*, 2025a.

Yucen Lily Li, Daohan Lu, Polina Kirichenko, Shikai Qiu, Tim G. J. Rudner, C. Bayan Bruss, and Andrew Gordon Wilson. Position: Supervised Classifiers Answer the Wrong Questions for OOD Detection. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 81689–81713, 2025b.

Jiachen Liang, RuiBing Hou, Minyang Hu, Hong Chang, Shiguang Shan, and Xilin Chen. Revisiting logit distributions for reliable out-of-distribution detection. In *Proc. of NeurIPS*, 2025.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *Proc. of ICLR*, 2018.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proc. of ICLR*, 2023.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow Matching Guide and Code. *ArXiv preprint abs/2412.06264*, 2024.

Chenhao Liu, Siyang Li, Luofei Tan, and Dongrui Wu. Temporal Out-of-Distribution Detection for Asynchronous Motor Imagery Brain-Computer Interfaces. *ArXiv preprint abs/2605.01014*, 2026.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proc. of NeurIPS*, 2020.

Silvia Isabel Lopez de Diego. Automated Interpretation of Abnormal Adult Electroencephalograms, 2017.

S. López, G. Suarez, D. Jungreis, I. Obeid, and J. Picone. Automated Identification of Abnormal Adult EEGs. *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2015. ISSN 2372-7241. doi: 10.1109/SPMB.2015.7405423.

Warren R. Morningstar, Cusuh Ham, Andrew G. Gallagher, Balaji Lakshminarayanan, Alexander A. Alemi, and Joshua V. Dillon. Density of states estimation for out of distribution detection. In *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3232–3240, 2021.

Merlijn Quincent Mulder, Matias Valdenegro-Toro, Andreea Ioana Sburlea, and Ivo Pascal de Jong. The Challenge of Out-Of-Distribution Detection in Motor Imagery BCIs. *ArXiv preprint abs/2603.13324*, 2026.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. *ArXiv preprint abs/1906.02994*, 2019a.

Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *Proc. of ICLR*, 2019b.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 427–436, 2015. doi: 10.1109/CVPR.2015.7298640.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience*, 10, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00196.

Devon I. Rubin. *Clinical Neurophysiology*. Contemporary Neurology Ser. 5th ed edition, 2021. ISBN 978-0-19-006785-4 978-0-19-006787-8.

Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. In *Proc. of NeurIPS*, 2020.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping*, 38(11):5391–5420, 2017. ISSN 1097-0193. doi: 10.1002/hbm.23730. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23730`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.23730.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *Proc. of ICLR*, 2020.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. EEG conformer: convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, PP, December 2022. doi: 10.1109/TNSRE.2022.3230250. URL `http://dx.doi.org/10.1109/{TNSRE}.2022.3230250`.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot Mueller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI Competition IV. *Frontiers in Neuroscience*, 6, 2012. ISSN 1662-453X. doi: 10.3389/fnins.2012.00055.

Mats Tveter, Thomas Tveitstøl, Christoffer Hatlestad-Hall, Hugo L. Hammer, and Ira R. J. Hebold Haraldsen. Uncertainty in deep learning for EEG under dataset shifts. *Artificial Intelligence in Medicine*, 174:103374, 2026. ISSN 0933-3657. doi: 10.1016/j.artmed.2026.103374.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. In *Proc. of NeurIPS*, 2022.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized Out-of-Distribution Detection: A Survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. ISSN 1573-1405. doi: 10.1007/s11263-024-02117-4.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection. *ArXiv preprint abs/2306.09301*, 2023.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. CTNet: a convolutional transformer network for EEG-based motor imagery classification. *Scientific Reports*, 14(1): 20237, August 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-71118-7. URL https://www.nature.com/articles/s41598-024-71118-7.

A OOD DETECTION RESULTS ON TUAB WITH CONFIDENCE INTERVALS

Table 3: **OOD detection performance on TUAB.** Performance is reported in terms of AUROC and bold values highlight the best result for each perturbation and severity level. The values in brackets denote confidence intervals computed through the percentile bootstrap procedure with 10,000 iterations. In each iteration, data was sampled with replacement while preserving the original ratio of ID to OOD samples. Discriminative-model-based methods (MSP, Energy, ODIN, ASH) perform largely at chance level, whereas generative-model-based methods achieve considerably better OOD detection with increasing performance for higher severity levels. SITN consistently achieves the best performance.

Perturbation	Severity	MSP	Energy	ODIN	ASH	LL	Typ.	DoSE	SITN
Resampling	125 Hz	0.517 [.515, .519]	0.500 [.498, .502]	0.514 [.513, .516]	0.533 [.531, .535]	0.369 [.367, .370]	0.559 [.557, .561]	0.556 [.554, .557]	**0.947** [.946, .948]
	150 Hz	0.457 [.455, .459]	0.479 [.477, .481]	0.453 [.451, .455]	0.534 [.532, .536]	0.275 [.274, .277]	0.646 [.644, .648]	0.647 [.645, .648]	**0.981** [.980, .982]
	200 Hz	0.383 [.381, .385]	0.451 [.449, .453]	0.377 [.375, .379]	0.522 [.520, .524]	0.143 [.142, .145]	0.816 [.814, .817]	0.826 [.825, .827]	**0.989** [.988, .989]
Channel Shuffle	10%	0.549 [.547, .551]	0.494 [.492, .496]	0.549 [.547, .551]	0.513 [.511, .515]	0.607 [.606, .609]	0.521 [.519, .523]	0.543 [.541, .545]	**0.754** [.753, .756]
	50%	0.553 [.551, .555]	0.431 [.429, .433]	0.555 [.554, .557]	0.494 [.492, .496]	0.793 [.792, .795]	0.682 [.681, .684]	0.711 [.709, .713]	**0.924** [.923, .925]
	100%	0.526 [.524, .528]	0.391 [.390, .393]	0.529 [.527, .531]	0.460 [.458, .462]	0.823 [.822, .825]	0.726 [.725, .728]	0.748 [.747, .750]	**0.942** [.941, .943]
Re-referencing	Cz	0.517 [.516, .519]	0.435 [.433, .436]	0.520 [.518, .522]	0.597 [.595, .598]	0.950 [.949, .950]	0.937 [.936, .938]	0.987 [.987, .987]	**0.998** [.997, .998]
	T7, T8	0.587 [.586, .589]	0.513 [.511, .514]	0.591 [.589, .593]	0.568 [.566, .570]	0.927 [.926, .927]	0.896 [.895, .897]	0.973 [.972, .973]	**0.997** [.997, .997]
	bipolar	0.752 [.750, .753]	0.621 [.619, .622]	0.757 [.755, .758]	0.719 [.718, .721]	0.940 [.939, .941]	0.919 [.918, .920]	0.962 [.962, .963]	**0.996** [.996, .996]
High-pass Filter	2 Hz	0.514 [.512, .516]	0.558 [.556, .560]	0.511 [.509, .513]	0.506 [.504, .508]	0.475 [.473, .477]	0.504 [.502, .506]	0.503 [.501, .505]	**0.604** [.602, .606]
	4 Hz	0.531 [.529, .533]	0.603 [.602, .605]	0.526 [.525, .528]	0.550 [.548, .551]	0.445 [.443, .447]	0.510 [.509, .512]	0.510 [.508, .511]	**0.749** [.747, .751]
	8 Hz	0.446 [.444, .448]	0.684 [.683, .686]	0.439 [.437, .441]	0.637 [.635, .639]	0.367 [.365, .369]	0.546 [.545, .548]	0.544 [.542, .545]	**0.872** [.870, .873]
Low-pass Filter	45 Hz	0.499 [.497, .501]	0.502 [.500, .504]	0.499 [.497, .501]	0.503 [.501, .504]	0.461 [.459, .463]	0.507 [.505, .509]	0.505 [.503, .507]	**0.822** [.821, .824]
	30 Hz	0.500 [.498, .502]	0.511 [.510, .513]	0.498 [.496, .500]	0.515 [.514, .517]	0.246 [.245, .248]	0.649 [.647, .651]	0.644 [.643, .646]	**0.986** [.985, .986]
	15 Hz	0.511 [.510, .513]	0.536 [.534, .538]	0.506 [.504, .508]	0.560 [.558, .562]	0.016 [.016, .016]	0.975 [.975, .976]	0.974 [.974, .975]	**0.990** [.990, .991]

B OOD DETECTION RESULTS ON CAUEEG WITH CONFIDENCE INTERVALS

Table 4: **OOD detection performance in terms of AUROC across different perturbations and severity levels on CAUEEG.** Bold values indicate the best performing metric for a given perturbation and severity. The values in brackets denote confidence intervals computed through the percentile bootstrap procedure with 10,000 iterations. In each iteration, data was sampled with replacement while preserving the original ratio of ID to OOD samples. Discriminative-model-based methods (MSP, Energy, ODIN, ASH) perform largely at chance level, whereas generative-model-based methods achieve considerably better OOD detection with increasing performance for higher severity levels.

Perturbation	Severity	MSP	Energy	ODIN	ASH	LL	Typ.	DoSE	SITN
Resampling	125 Hz	0.531 [.527, .536]	0.517 [.513, .522]	0.540 [.536, .544]	0.540 [.536, .545]	0.286 [.282, .290]	0.587 [.583, .592]	0.760 [.756, .763]	**0.956** [.955, .958]
	150 Hz	0.497 [.492, .501]	0.520 [.516, .524]	0.514 [.510, .519]	0.590 [.586, .595]	0.166 [.163, .169]	0.732 [.728, .735]	0.886 [.883, .889]	**0.984** [.982, .985]
	200 Hz	0.470 [.465, .474]	0.531 [.527, .535]	0.495 [.491, .499]	0.658 [.654, .662]	0.056 [.055, .058]	0.906 [.903, .908]	0.965 [.963, .966]	**0.992** [.991, .993]
Channel Shuffle	10%	0.496 [.492, .500]	0.502 [.498, .506]	0.505 [.500, .509]	0.518 [.514, .522]	0.622 [.618, .626]	0.552 [.548, .556]	0.711 [.708, .715]	**0.759** [.755, .763]
	50%	0.443 [.439, .447]	0.490 [.485, .494]	0.470 [.466, .474]	0.570 [.566, .574]	0.841 [.838, .844]	0.766 [.763, .770]	**0.968** [.966, .969]	0.956 [.955, .958]
	100%	0.408 [.404, .412]	0.483 [.479, .488]	0.436 [.432, .441]	0.583 [.579, .587]	0.879 [.877, .882]	0.819 [.816, .822]	**0.982** [.981, .983]	0.968 [.966, .969]
Re-referencing	Cz	0.281 [.277, .285]	0.230 [.227, .233]	0.305 [.301, .309]	0.366 [.361, .370]	0.992 [.991, .993]	0.991 [.990, .991]	**1.000** [.999, 1.000]	0.997 [.996, .997]
	T7, T8	0.305 [.301, .309]	0.361 [.357, .365]	0.339 [.335, .343]	0.471 [.466, .475]	0.982 [.981, .983]	0.976 [.975, .977]	**1.000** [.999, 1.000]	0.997 [.997, .997]
	bipolar	0.426 [.422, .431]	0.439 [.434, .443]	0.454 [.450, .458]	0.562 [.557, .566]	0.990 [.990, .991]	0.989 [.988, .990]	**1.000** [.999, 1.000]	0.997 [.996, .997]
High-pass Filter	2 Hz	0.520 [.515, .524]	**0.640** [.636, .644]	0.512 [.507, .516]	0.532 [.527, .536]	0.454 [.450, .459]	0.507 [.503, .512]	0.566 [.562, .570]	0.606 [.601, .610]
	4 Hz	0.507 [.503, .512]	0.685 [.681, .689]	0.494 [.490, .498]	0.578 [.573, .582]	0.408 [.403, .412]	0.518 [.513, .522]	0.594 [.590, .598]	**0.756** [.753, .760]
	8 Hz	0.518 [.514, .522]	0.792 [.789, .796]	0.494 [.490, .499]	0.721 [.717, .724]	0.291 [.287, .295]	0.589 [.585, .593]	0.626 [.622, .630]	**0.899** [.896, .901]
Low-pass Filter	45 Hz	0.496 [.492, .500]	0.501 [.496, .505]	0.496 [.491, .500]	0.502 [.498, .506]	0.446 [.442, .450]	0.505 [.500, .509]	0.544 [.539, .548]	**0.822** [.819, .825]
	30 Hz	0.495 [.491, .499]	0.510 [.505, .514]	0.492 [.488, .497]	0.515 [.511, .519]	0.186 [.182, .189]	0.700 [.696, .703]	0.943 [.941, .945]	**0.987** [.986, .988]
	15 Hz	0.557 [.553, .562]	0.564 [.560, .568]	0.556 [.552, .560]	0.563 [.558, .567]	0.007 [.006, .008]	0.989 [.988, .989]	**0.998** [.998, .998]	0.994 [.993, .995]

C MAHALANOBIS DISTANCE

To compute the Mahalanobis distance (Lee et al., 2018), we first extracted the penultimate-layer feature representations from the pre-trained discriminative model, taking the activations immediately following the global average pooling layer. Using the in-distribution validation set, we estimated the class-conditional mean feature vectors alongside a single, shared covariance matrix tied across all classes. To ensure numerical stability, a small regularisation term ($\epsilon = 10^{-5}$) was added to the diagonal of the shared covariance matrix prior to computing its inverse (the shared precision matrix). During evaluation, the Mahalanobis distance was calculated between each test sample's feature vector and each of the class-conditional means using this shared precision matrix. A sample's final score was then defined as its minimum Mahalanobis distance across all available classes. OOD samples are expected to have higher scores indicating that their embeddings are situated further from the known in-distribution classes.

OOD detection results are shown in Table 5. In contrast to other discriminative methods, the Mahalanobis distance yielded superior OOD detection performance under the Resampling, Channel Shuffle, and Re-referencing perturbations. However, it exhibited a severe vulnerability to the High-pass Filter perturbation, where it performed worse than random chance.

Table 5: **Mahalanobis distance OOD detection performance in terms of AUROC across different perturbations and severity levels.** Compared to other discriminative approaches, Mahalanobis distances show better performance for the Resampling, Channel Shuffle, and Re-referencing perturbations, but also a strong failure for the High-pass Filter perturbation, where they consistently underperform even random guessing.

Perturbation	Severity	TUAB	CAUEEG
Resampling	125 Hz	0.548	0.544
	150 Hz	0.594	0.586
	200 Hz	0.654	0.640
Channel Shuffle	10%	0.583	0.604
	50%	0.741	0.763
	100%	0.775	0.799
Re-referencing	Cz	0.851	0.925
	T7, T8	0.794	0.845
	bipolar	0.786	0.861
High-pass Filter	2 Hz	0.348	0.263
	4 Hz	0.321	0.252
	8 Hz	0.285	0.153
Low-pass Filter	45 Hz	0.500	0.501
	30 Hz	0.506	0.505
	15 Hz	0.541	0.519

D LOGITGAP

Additional results for the recent LogitGap method (Liang et al., 2025), a discriminative approach based on the logit distribution, are shown in Table 6. Results are consistent with those of other discriminative approaches.

Table 6: **LogitGap OOD detection performance in terms of AUROC across different perturbations and severity levels.** Results are similar to those of other discriminative approaches in Table 1.

Perturbation	Severity	TUAB	CAUEEG
Resampling	125 Hz	0.488	0.540
	150 Hz	0.418	0.516
	200 Hz	0.350	0.498
Channel Shuffle	10%	0.541	0.505
	50%	0.581	0.471
	100%	0.576	0.437
Re-referencing	Cz	0.504	0.304
	T7, T8	0.593	0.340
	bipolar	0.779	0.454
High-pass Filter	2 Hz	0.530	0.516
	4 Hz	0.569	0.498
	8 Hz	0.544	0.500
Low-pass Filter	45 Hz	0.495	0.496
	30 Hz	0.486	0.492
	15 Hz	0.477	0.556

E SAMPLE SIZE DEPENDENCE

A potential limitation of generative methods could be a larger sample size requirement. To investigate this possibility, we performed a data scaling analysis where the amount of training data was varied while the validation and test data was kept the same. Starting from a subset of just 25 participants and 5 segments per participant, we increased the sample size in two different ways: either increasing the number of participants or the number of segments per participant. These experiments were conducted on the TUAB dataset and subsampling of participants was stratified by the normality label. We used the resampling perturbation at 200 Hz for OOD detection evaluations.

As shown in Figure 4, OOD detection performance generally increased with the amount of training data, where a larger number of participants (solid lines) was more impactful than an equal number of segments from a smaller participant cohort (dashed lines). Generative methods (SITN in particular) performed surprisingly well even with very limited data and dominated discriminative methods across all evaluated sample sizes. The failure mode for log-likelihoods where performance decreased with more training data is discussed in Appendix F.

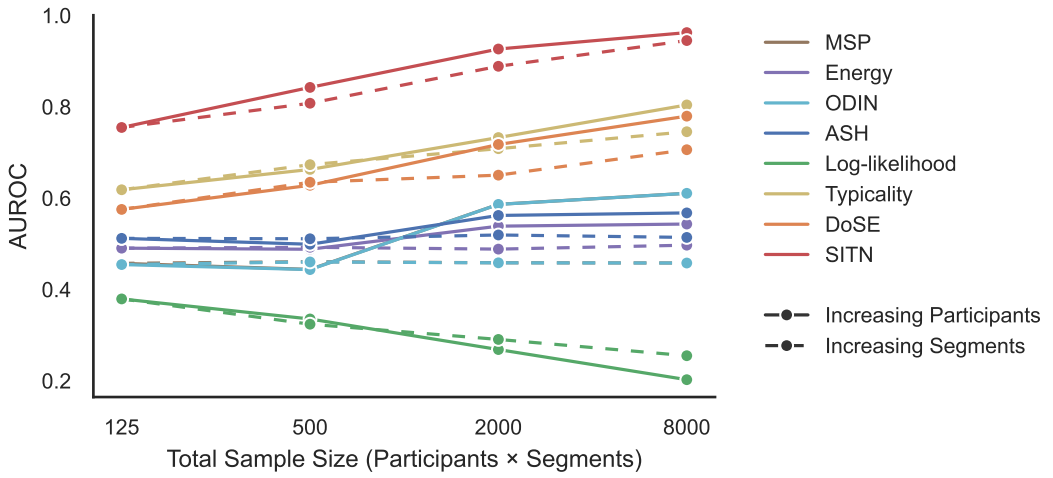

Figure 4: **Sample size dependence of OOD detection performance.** Starting from a baseline of just 25 participants and 5 segments per participant (= 125 samples in total), the training data was gradually increased by either increasing the number of participants (solid lines) or segments per participant (dashed lines). The sampling rate perturbation at 200 Hz was used to generate OOD samples. OOD detection performance generally increased with the amount of training data (see Appendix F for a discussion of the likelihood). Even in the smallest data regimes, generative methods dominated discriminative approaches. Note that MSP is occluded by ODIN, which showed almost identical performance across sample sizes.

F LIMITATIONS OF LIKELIHOODS FOR OOD DETECTION

Several works in the computer vision literature have identified critical limitations of likelihoods for OOD detection (Nalisnick et al., 2019b; Serrà et al., 2020; Schirrmeister et al., 2020; Kirichenko et al., 2020; Havtorn et al., 2021; Kamkari et al., 2024), which motivated the development of methods like Typicality, DoSE, and SITN. In widely replicated experiments, it was shown that generative models trained on CIFAR-10 consistently assign higher likelihoods to samples from other datasets like MNIST or SVHN (Nalisnick et al., 2019b; Serrà et al., 2020; Schirrmeister et al., 2020; Kirichenko et al., 2020). Investigating this issue, Serrà et al. (2020) further observed a strong correlation between likelihood and sample complexity (measured by PNG compression size), showing that "simpler" samples like constant-colour images were consistently assigned higher likelihoods than more "complex" ones.

We investigated and were able to replicate the same behaviour for EEG data. Panel A in Figure 5 shows the log-likelihood distribution across ID test samples (100 Hz) compared to the distributions for the corresponding OOD samples obtained through the resampling perturbation at different severities. It can be seen that OOD samples were consistently assigned higher likelihoods, with the highest likelihoods assigned to samples perturbed at the highest severity level. As mentioned in Section 3.1 and illustrated in Figure 2, samples perturbed with this perturbation appear increasingly smoothed with higher severities. Furthermore, Panel B in the same figure shows the (ID) EEG segments with the lowest and highest likelihood in the dataset. Consistent with the complexity observations by Serrà et al. (2020), the highest likelihoods are assigned to completely flat EEG segments, whereas high-amplitude artefacts resulted in particularly low likelihoods.

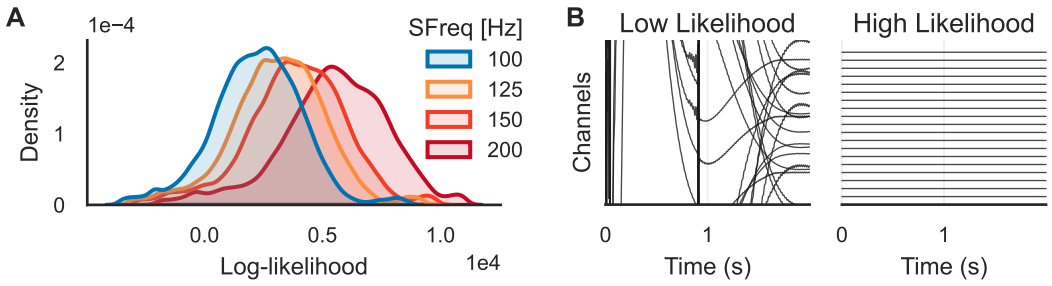

Figure 5: **Likelihood limitations. A** Distribution of log-likelihoods across ID samples (100 Hz) and OOD samples obtained with the resampling perturbation at increasing severities of 125, 150, and 200 Hz. As previously observed in the context of computer vision datasets, OOD samples are consistently assigned higher likelihoods than the in-distribution data. **B** EEG segments that were assigned the lowest and highest likelihoods, respectively. As in computer vision, where "simpler" images like constant-colour images are assigned high likelihoods, EEG segments with completely flat channels also result in high likelihoods. High-amplitude artefacts result in the opposite extreme and are assigned very small likelihoods.

G TRAINING DETAILS

For the generative model, the UNet architecture following Ho et al. (2020), as implemented in Dhariwal & Nichol (2021) was used with the hyperparameters detailed in Table 7. Following the flow matching objective of Lipman et al. (2023), we constructed a linear interpolant between a Gaussian noise sample $Z \sim \mathcal{N}(0, I)$ and a data sample $X \sim P_{data}$: $X_t = (1-t)Z + tX, \quad t \sim \mathcal{U}[0,1]$ and minimised the mean-squared error between the network output and the target vector field $X - Z$.

For the discriminative model, a Temporal Convolutional Network (TCN) (Bai et al., 2018) was used with the hyperparameters detailed in Table 8. This model was trained using a supervised cross-entropy loss for the respective prediction task.

Both models were optimised with AdamW (learning rate 10^{-4}, weight decay 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) and global gradient norm clipping (max norm 1.0). The generative and discriminative models were trained for 1,000,000 and 50,000 gradient steps, respectively, at a batch size of 128 and with weights restored to the checkpoint of lowest validation loss. It was empirically verified that the validation loss had converged within the maximum number of gradient steps for both models.

Table 7: UNet hyperparameters.

Hyperparameter	Value
Channels	64
Depth	2
Channel multipliers	1, 2, 2, 2
Heads	1
Attention resolution	8
Dropout	0.0

Table 8: TCN hyperparameters.

Hyperparameter	Value
Blocks	4
Filters	64
Kernel size	5
Dilations	1, 2, 4, 8
Dropout	0.0

H HYPERPARAMETER SENSITIVITY ANALYSIS

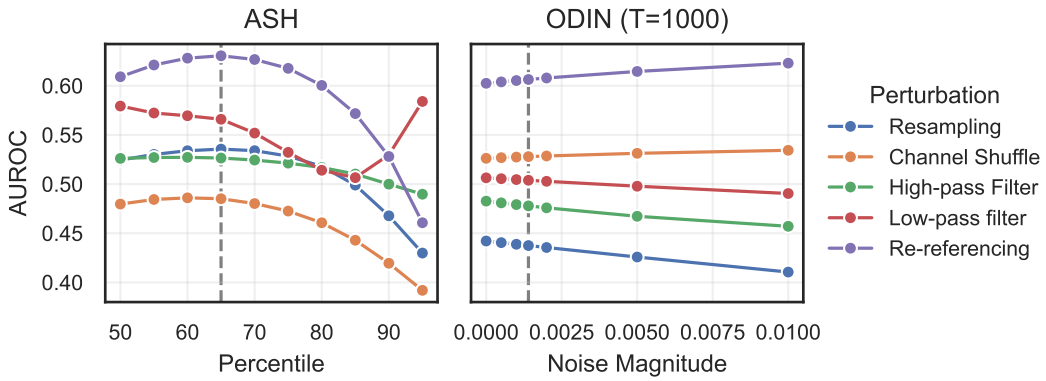

Figure 6: **Hyperparameter Sensitivity Analysis.** The plots show OOD detection AUROC values for each perturbation (averaged across severity levels) for ASH (left) and ODIN (right) across varying hyperparameter settings. Performance was evaluated on the TUAB validation split. The dashed lines indicate the default hyperparameters used in the main experiments (65 for ASH, 0.0014 for ODIN).

I Model Architecture Control Experiments

Figure 7 shows a comparison of different model architectures (ShallowFBCSPNet (Schirrmeister et al., 2017), EEGNet (Lawhern et al., 2018), EEGConformer (Song et al., 2022), CTNet (Zhao et al., 2024), ATCNet (Altaheri et al., 2023)) to the TCN model used for the discriminative methods in the main experiments. We used the model implementations from the braindecode library (Aristimunha et al., 2025) with default hyperparameters. The first row shows AUROC values for model performance in the downstream normality prediction task on TUAB (evaluated on in-distribution test data) while the rest of the rows show AUROC values for OOD detection performance across the Resampling and Channel Shuffle perturbations across the different OOD detection methods on TUAB.

TCN and ShallowFBCSPNet achieved the highest performance on the downstream prediction task with EEGConformer close behind. OOD detection performance varied somewhat across models and perturbations, but is qualitatively similar to the results in Table 1 with clearly inferior performance across all model architectures compared to generative approaches.

			Shallow	EEGNet	EEGConf	CTNet	ATCNet	TCN
Downstream (ID) performance (AUROC)			0.872	0.834	0.860	0.841	0.814	0.873
Resampling	125 Hz	MSP	0.578	0.596	0.571	0.576	0.599	0.517
Resampling	125 Hz	energy	0.473	0.622	0.601	0.434	0.573	0.500
Resampling	125 Hz	ODIN	0.577	0.596	0.570	0.576	0.599	0.514
Resampling	125 Hz	ASH	0.442	0.593	0.495	0.465	0.451	0.533
Resampling	150 Hz	MSP	0.528	0.626	0.545	0.593	0.653	0.457
Resampling	150 Hz	energy	0.370	0.679	0.678	0.449	0.593	0.479
Resampling	150 Hz	ODIN	0.527	0.626	0.543	0.593	0.653	0.453
Resampling	150 Hz	ASH	0.403	0.629	0.549	0.492	0.410	0.534
Resampling	200 Hz	MSP	0.419	0.654	0.481	0.624	0.684	0.383
Resampling	200 Hz	energy	0.245	0.720	0.740	0.541	0.589	0.451
Resampling	200 Hz	ODIN	0.418	0.655	0.478	0.623	0.685	0.377
Resampling	200 Hz	ASH	0.362	0.652	0.621	0.601	0.386	0.522
Channel Shuffle	10%	MSP	0.591	0.523	0.544	0.525	0.530	0.549
Channel Shuffle	10%	energy	0.541	0.510	0.531	0.510	0.516	0.494
Channel Shuffle	10%	ODIN	0.594	0.523	0.544	0.525	0.530	0.549
Channel Shuffle	10%	ASH	0.488	0.510	0.485	0.526	0.486	0.513
Channel Shuffle	50%	MSP	0.687	0.577	0.560	0.593	0.622	0.553
Channel Shuffle	50%	energy	0.568	0.538	0.658	0.540	0.563	0.431
Channel Shuffle	50%	ODIN	0.694	0.577	0.559	0.593	0.622	0.555
Channel Shuffle	50%	ASH	0.470	0.538	0.539	0.600	0.437	0.494
Channel Shuffle	100%	MSP	0.684	0.594	0.531	0.612	0.662	0.526
Channel Shuffle	100%	energy	0.571	0.550	0.710	0.551	0.578	0.391
Channel Shuffle	100%	ODIN	0.691	0.594	0.530	0.612	0.663	0.529
Channel Shuffle	100%	ASH	0.466	0.551	0.588	0.620	0.406	0.460

Figure 7: **Model Architecture Comparison on TUAB.** The first row shows downstream performance on the normality prediction task whereas the rest of the rows show OOD detection performance across the Resampling and Channel Shuffle perturbations at different severities and across different OOD detection methods. Performance is reported in terms of AUROC scores.

J Exploratory Cross-dataset Experiments

We conducted a number of cross-dataset experiments where we evaluated the TUAB-trained methods on CAUEEG and conversely the CAUEEG-trained methods on TUAB. We consider this analysis exploratory as it is unclear to what degree samples from one of these datasets should be considered OOD for a model trained on the other dataset. Aside from potential (unquantified) hardware differences, participants in the TUAB dataset could in principle also be assigned a label from the dementia diagnosis task on CAUEEG, such that EEG recordings from one dataset should not necessarily be considered OOD for a model trained on the other one. Rather than treating this setup as a benchmarking task to evaluate OOD detection performance, we are interested to what degree cross-dataset samples result in high OOD scores and if there are differences across participant subgroups.

Table 9 shows AUROC values analogous to those reported for OOD detection performance. Overall, the different methods show moderate discrimination between the two datasets. Scores from DoSE and SITN, the two strongest OOD detection methods according to the results in Section 4.1, do not separate samples from the two datasets when trained on TUAB (AUROC < 0.55), but achieve AUROC values of 0.703 and 0.740 respectively when trained on CAUEEG. This would be consistent with CAUEEG forming a subset of the TUAB data distribution.

Table 9: **Cross-dataset discrimination across OOD detection methods.** Models were trained on one dataset and then evaluated using the test split of the training dataset as ID and the test split of the other dataset as OOD data (e.g. TUAB→CAUEEG denotes trained on TUAB and using CAUEEG as OOD). The degree to which OOD scores discriminate between samples from the two datasets is reported in terms of AUROC. Note that ID and OOD should be interpreted with a grain of salt in this setup as it is unclear to what degree cross-dataset samples are OOD.

Condition	MSP	Energy	ODIN	ASH	LL	Typ.	DoSE	SITN
TUAB → CAUEEG	0.594	0.478	0.598	0.517	**0.669**	0.432	0.462	0.544
CAUEEG → TUAB	0.489	0.558	0.477	0.514	0.398	0.643	0.703	**0.740**

In the following figures, we compare OOD scores across different subgroups within the target dataset. To do so, we aggregated OOD scores on the participant level by averaging across samples, resulting in an average OOD score per participant. Furthermore, to facilitate comparisons, we normalised OOD scores by applying min-max scaling to map the raw outputs of each OOD detection method to a standard $[0, 1]$ range. Additionally, scores were inverted where necessary, such that a higher normalised OOD score always indicates "more OOD".

Figure 8 shows normalised OOD scores across the different diagnosis labels on CAUEEG (normal, mild cognitive impairment, dementia) for the TUAB-trained models. Trends across diagnosis groups are inconsistent across methods and scores are strongly overlapping. Figure 9 shows the same scores across age, with the colours corresponding to the same diagnosis groups (normal in green, MCI in orange, dementia in blue). No clear trends are discernible, which might be expected given that the age range of the TUAB training data spans the full age range of the CAUEEG dataset. Figure 10 shows normalised scores for the CAUEEG-trained models on TUAB, stratified by the EEG normality label. Although scores are strongly overlapping for all methods, abnormal EEGs tended to be assigned somewhat higher OOD scores by SITN. Finally, Figure 11 shows the same OOD scores across age on TUAB. As for CAUEEG, no clear trend is discernible, although in this case one might have expected younger participants to look more OOD given that only 14 out of 950 participants in the CAUEEG training data (less than 2%) are aged younger than 50.

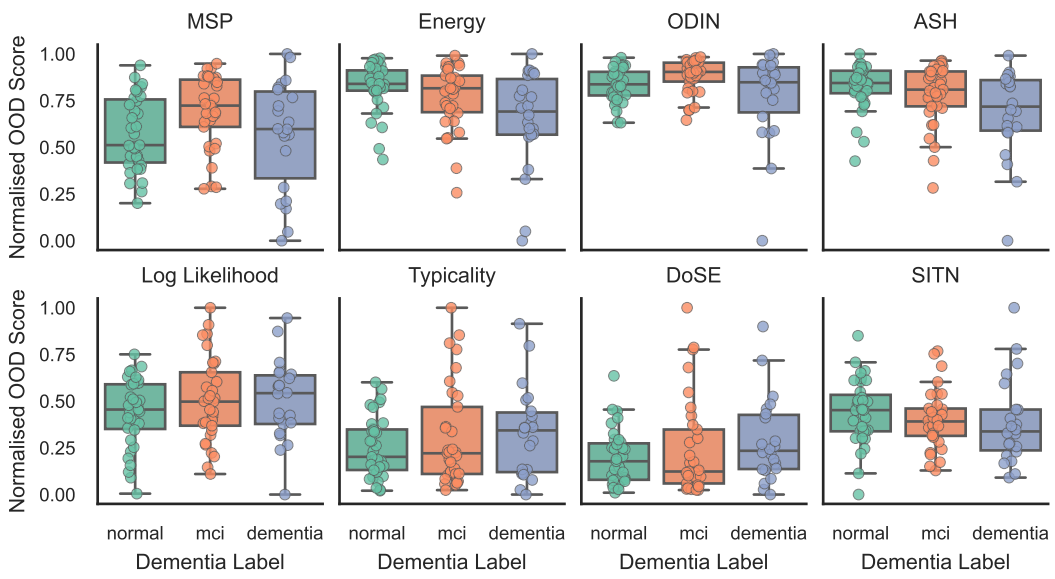

Figure 8: **Cross-dataset subgroup analysis of OOD scores across diagnosis labels on CAUEEG.** The results show normalised participant-level OOD scores for samples from the CAUEEG dataset assigned by methods trained on TUAB. Scores are largely overlapping without a consistent trend across diagnosis groups.

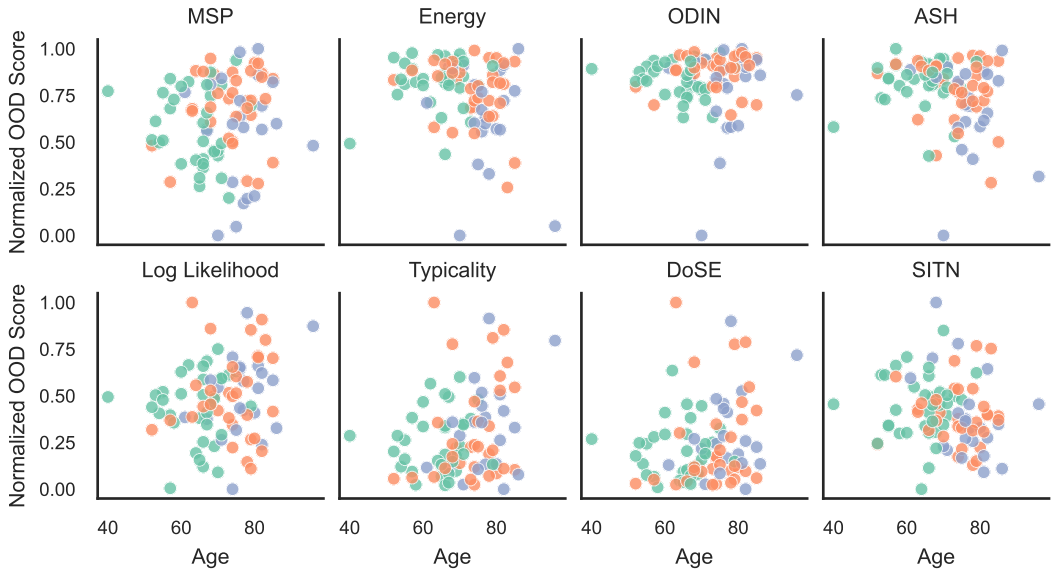

Figure 9: **Cross-dataset subgroup analysis of OOD scores across age on CAUEEG.** The results show normalised participant-level OOD scores for samples from the CAUEEG dataset assigned by methods trained on TUAB. No clear trend is observable.

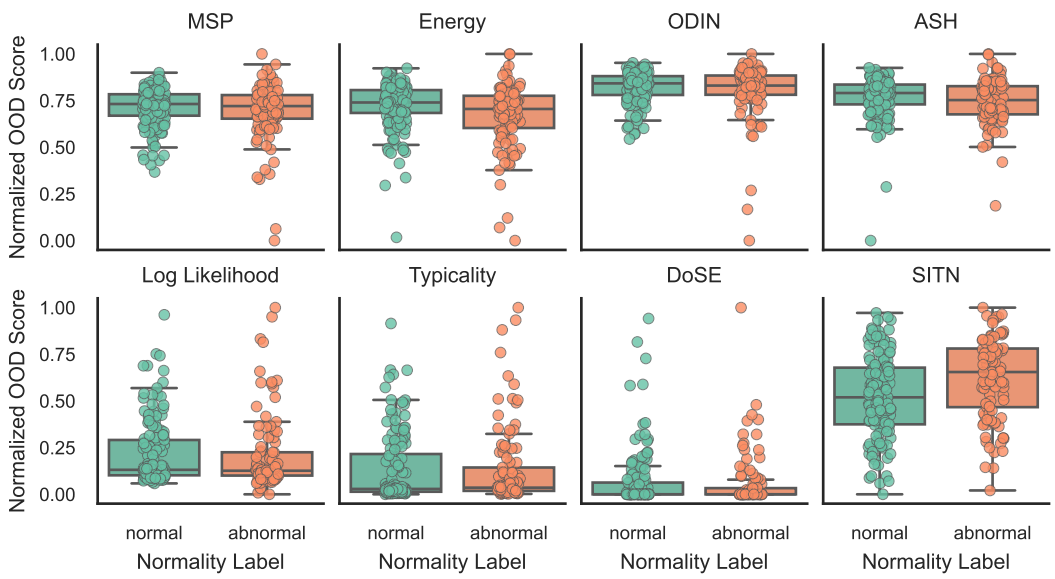

Figure 10: **Cross-dataset subgroup analysis of OOD scores across EEG normality labels on TUAB.** The results show normalised participant-level OOD scores for samples from the TUAB dataset assigned by methods trained on CAUEEG. Scores are largely overlapping, although SITN assigned somewhat higher OOD scores to abnormal EEG recordings.

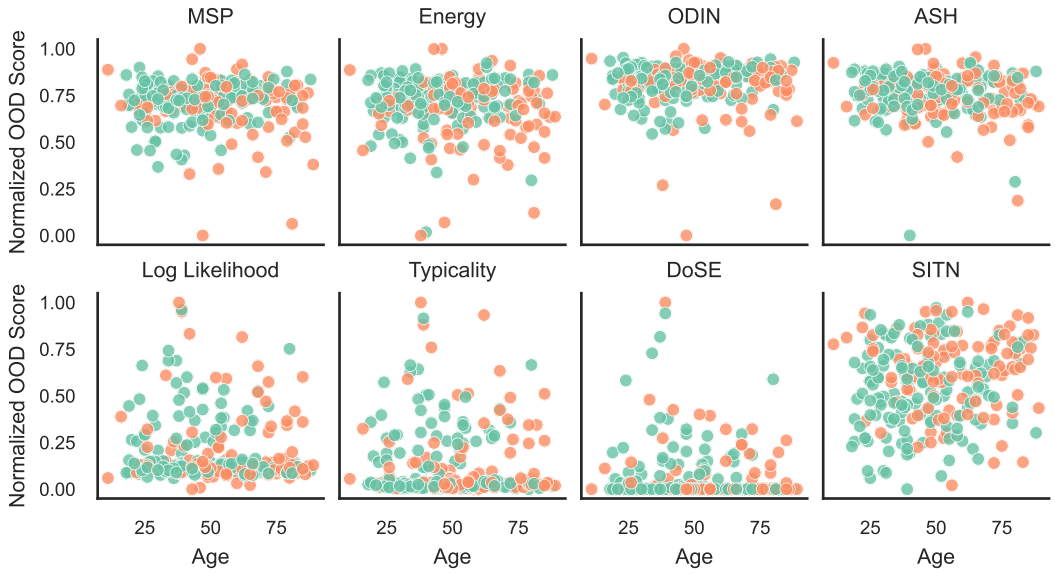

Figure 11: **Cross-dataset subgroup analysis of OOD scores across age on TUAB.** The results show normalised participant-level OOD scores for samples from the TUAB dataset assigned by methods trained on CAUEEG. No clear trends are discernible.

K CROSS-PARTICIPANT BCI EXPERIMENT

We evaluated the different OOD detection methods in a cross-participant setup on the BNCI 2014-001 Motor Imagery dataset (Tangermann et al., 2012). This dataset contains EEG data for a motor imagery task from nine participants. Two sessions (recorded on different days) are available for each participant, each of which in turn consists of six runs with 12 trials per class. In our setup, we only used trials for the left hand and right hand classes, resulting in 144 trials per session and participant. We used the official train-test split, which assigns the first session of each participant to train and the second to test. Additionally, we used the last run in the training session as validation data. For our cross-participant experiment, we trained the classification model and the OOD detection methods only on the training data of the first participant. We then evaluated the methods on the test data of the training participant as well as on the test data of all other participants.

Figure 12 shows the results with prediction accuracy for the motor imagery task and OOD scores aggregated to the participant level through averaging. For better visualisation, we flipped scores where necessary, such that higher scores on the x-axis always correspond to higher OOD scores. SITN is the only method to clearly separate the (ID) training participant from all (OOD) participants. On six out of eight of the unseen participants, prediction accuracy for the motor imagery task dropped considerably below the within-participant test performance. MSP assigned the lowest OOD scores to the two participants where model performance was higher than for the training participant, however, it also assigned the third lowest OOD score to a participant where performance dropped below chance level and which was flagged as highly OOD by all generative methods as well as the Energy score and ASH.

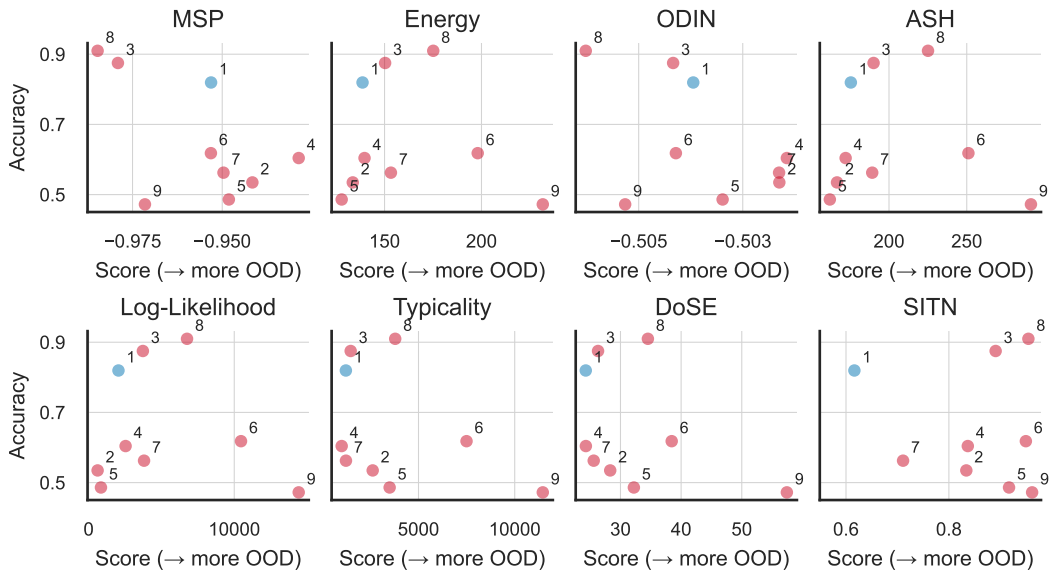

Figure 12: **Cross-participant experiments on the BNCI 2014-001 motor imagery dataset.** Numbers next to the points indicate participant id. Methods were trained on the training data from participant 1 and evaluated on the test data from all participants. OOD scores were aggregated on the participant-level through averaging and aligned such that higher scores always indicate more OOD. Only SITN clearly separates test data from the training participant (ID) from the other participants (OOD). MSP assigns the lowest OOD scores to two participants where model performance was even higher than for the training participant, but also the 3rd lowest score to the participant where model performance was lowest and which was flagged as highly OOD by most other methods.

L Full Downstream Performance Results

The following pages show the full results for the downstream performance evaluations across both tasks and datasets (normality prediction on TUAB (binary classification, chance level 50%); dementia diagnosis on CAUEEG (3-way classification into normal, MCI, dementia, chance level $\approx 33\%$)) for all perturbations and OOD detection methods. Performance is shown in terms of accuracy across percentile bins of the OOD score, where the x ticks indicate the upper edge of the corresponding bin (i.e. the first bin contains samples from the 0-0.1 percentile, the second one 0.1-1, etc.). The histograms indicate the number of ID and OOD samples in each bin. The arrows in the top left/right of each plot indicate the OOD direction for the corresponding metric. The captions for each figure indicate the task and dataset as well as the perturbation used to generate the OOD data.

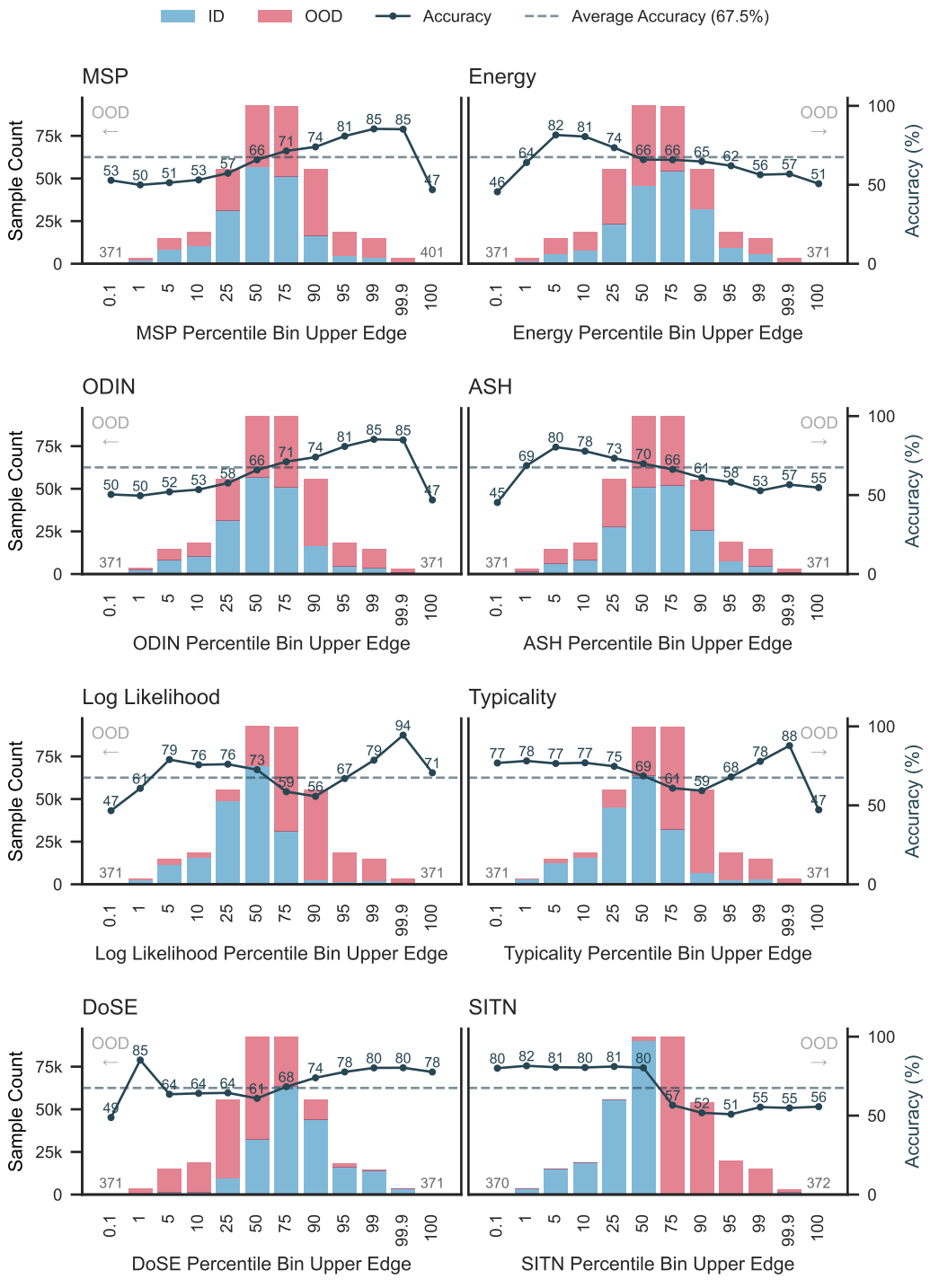

Figure 13: **Normality prediction on TUAB — resampling perturbation (200 Hz)**

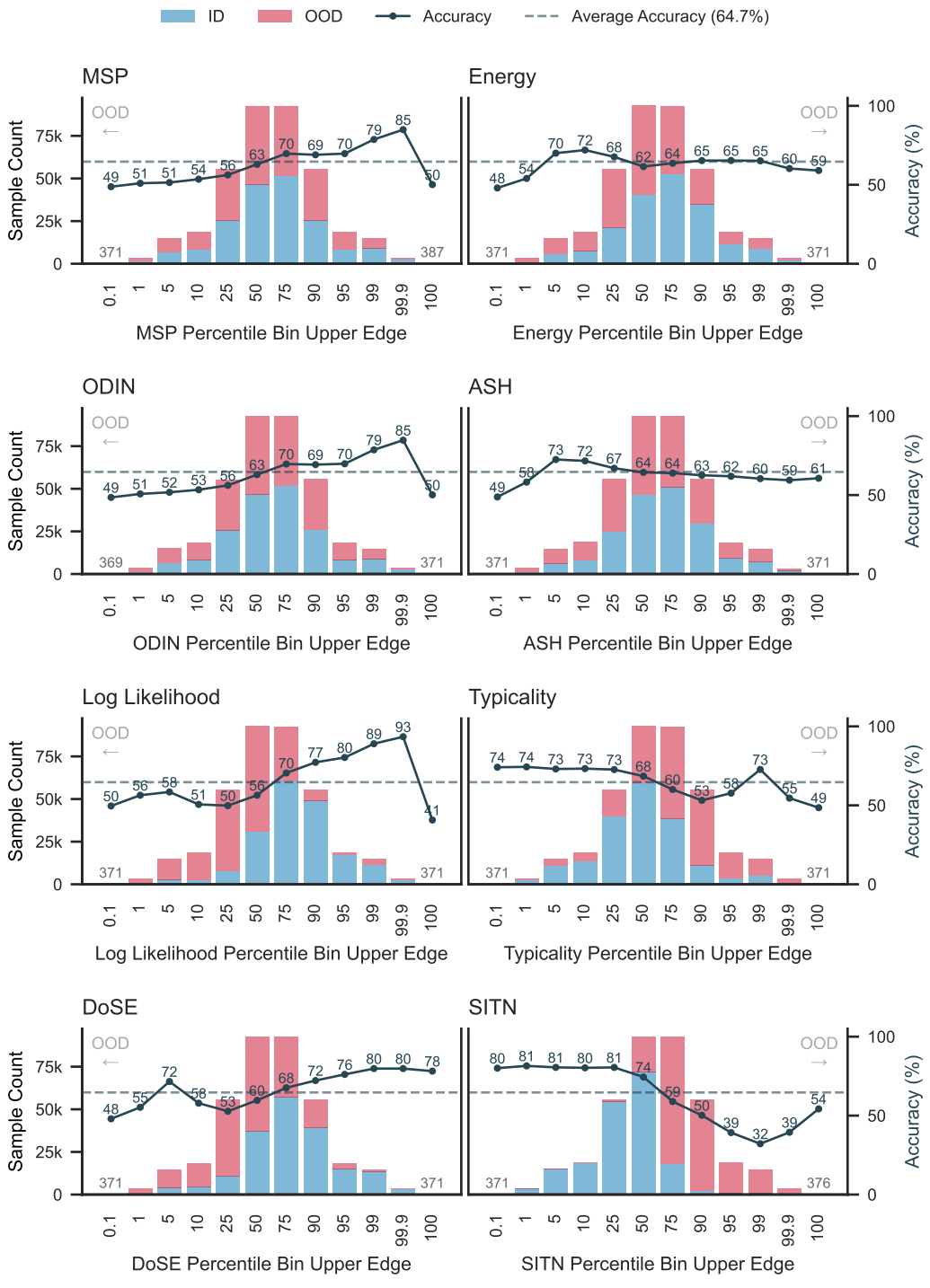

Figure 14: **Normality prediction on TUAB — channel shuffle perturbation (100%)**

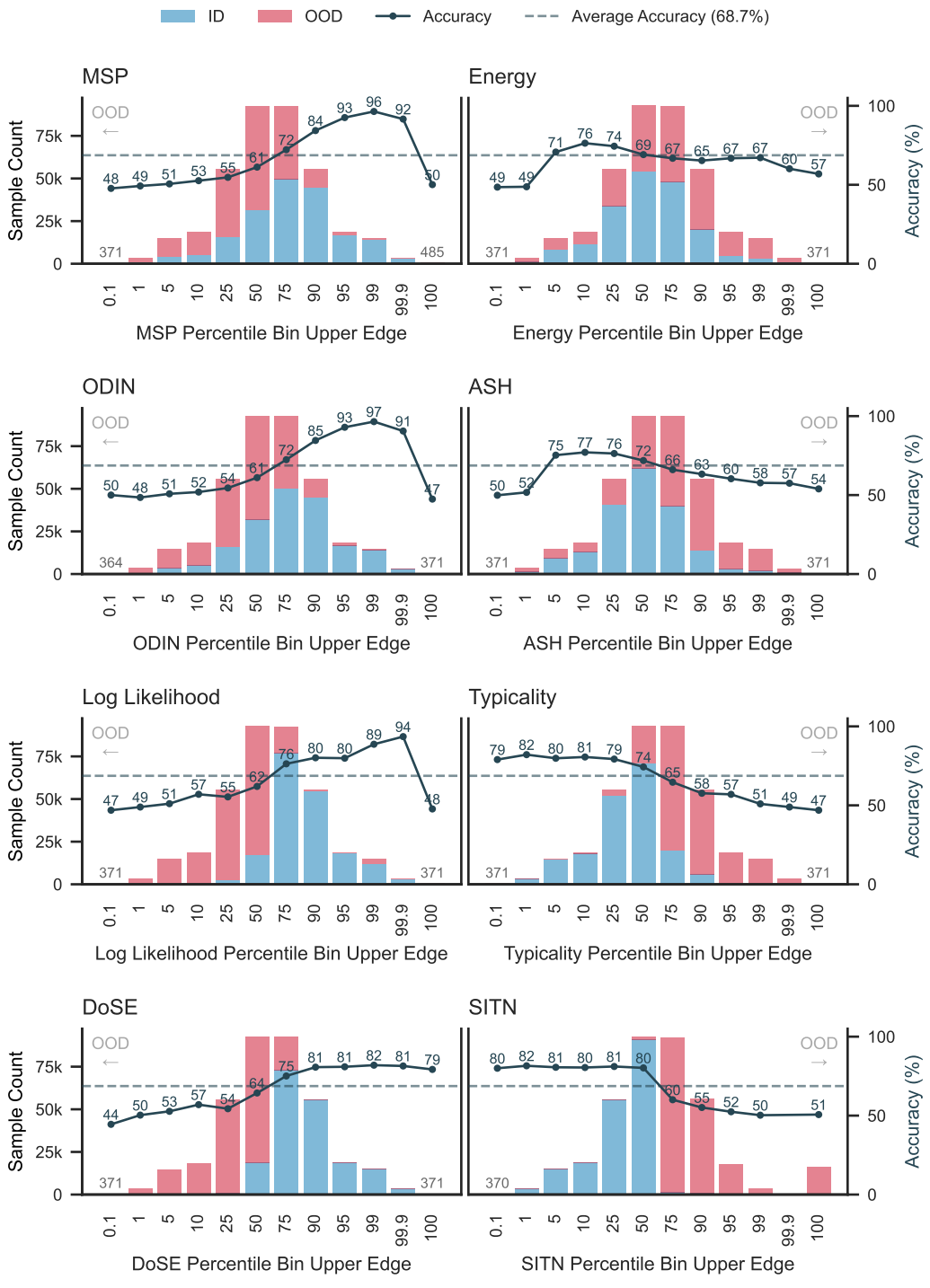

Figure 15: **Normality prediction on TUAB — re-referencing perturbation (bipolar)**
Note: The irregular bin sizes at the upper percentiles of the SITN plot are due to metric saturation, where a large number of tied maximum scores are grouped into the final bin.

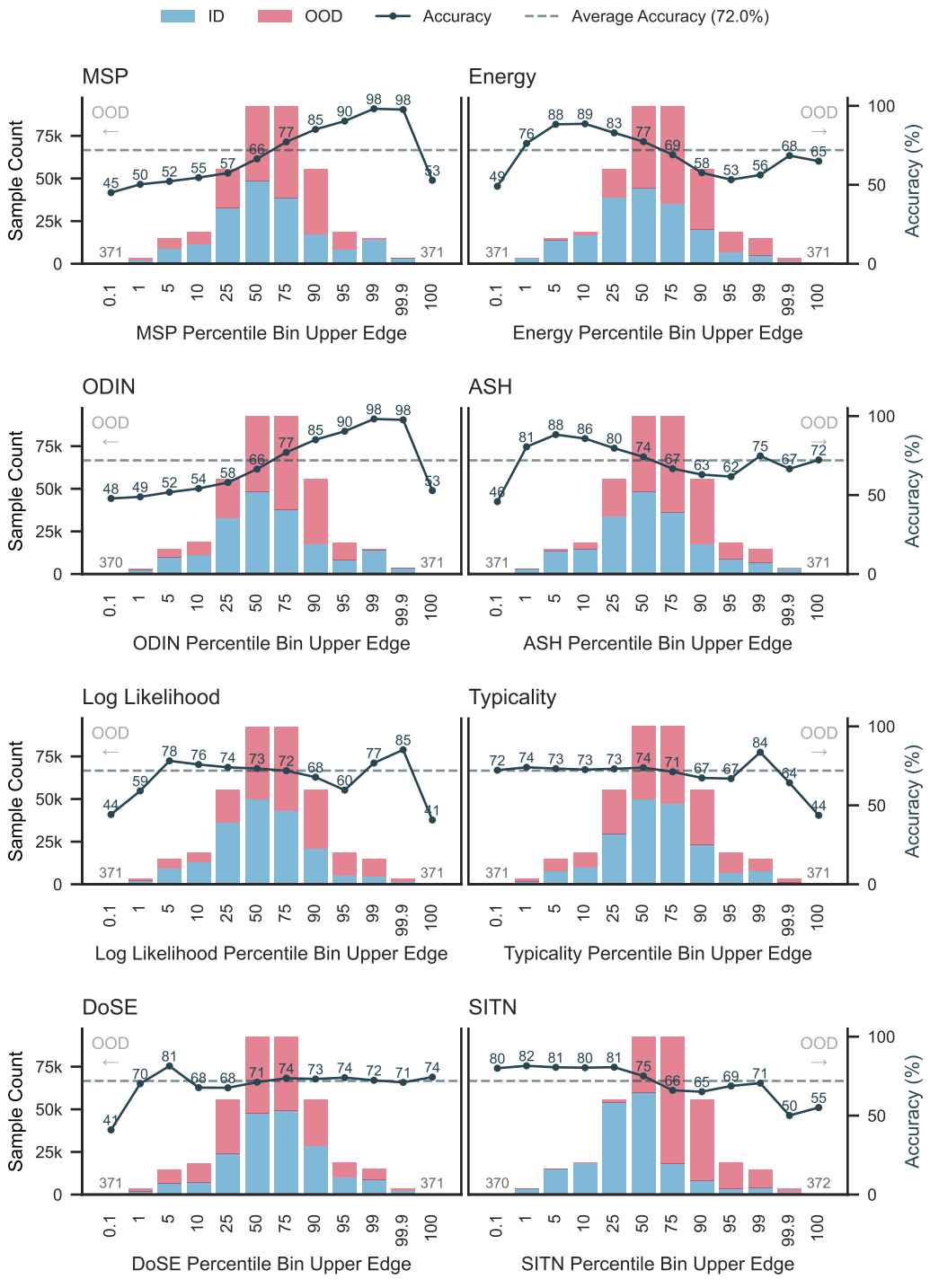

Figure 16: **Normality prediction on TUAB — high-pass filter perturbation (8 Hz)**

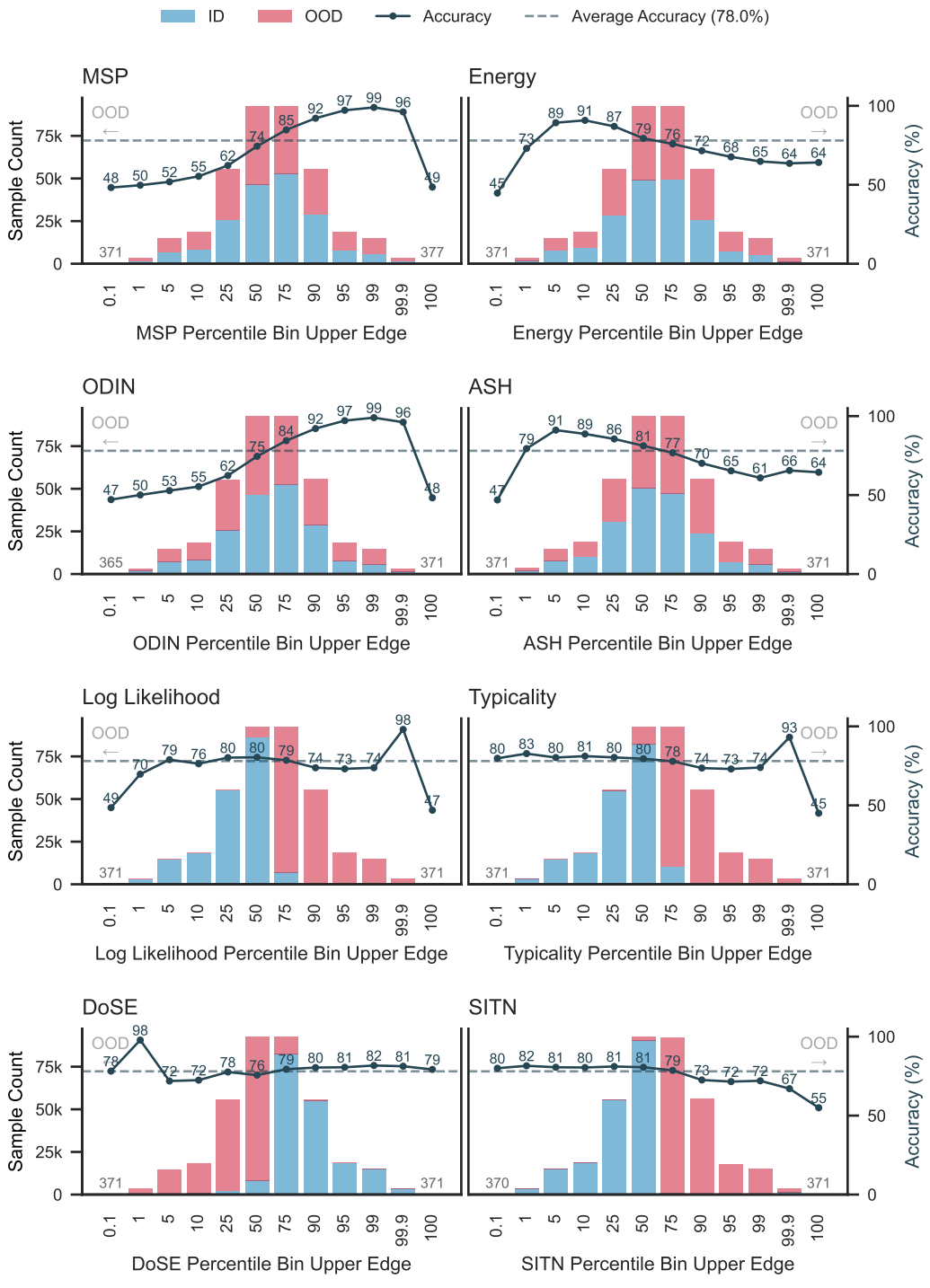

Figure 17: **Normality prediction on TUAB — low-pass filter perturbation (15 Hz)**

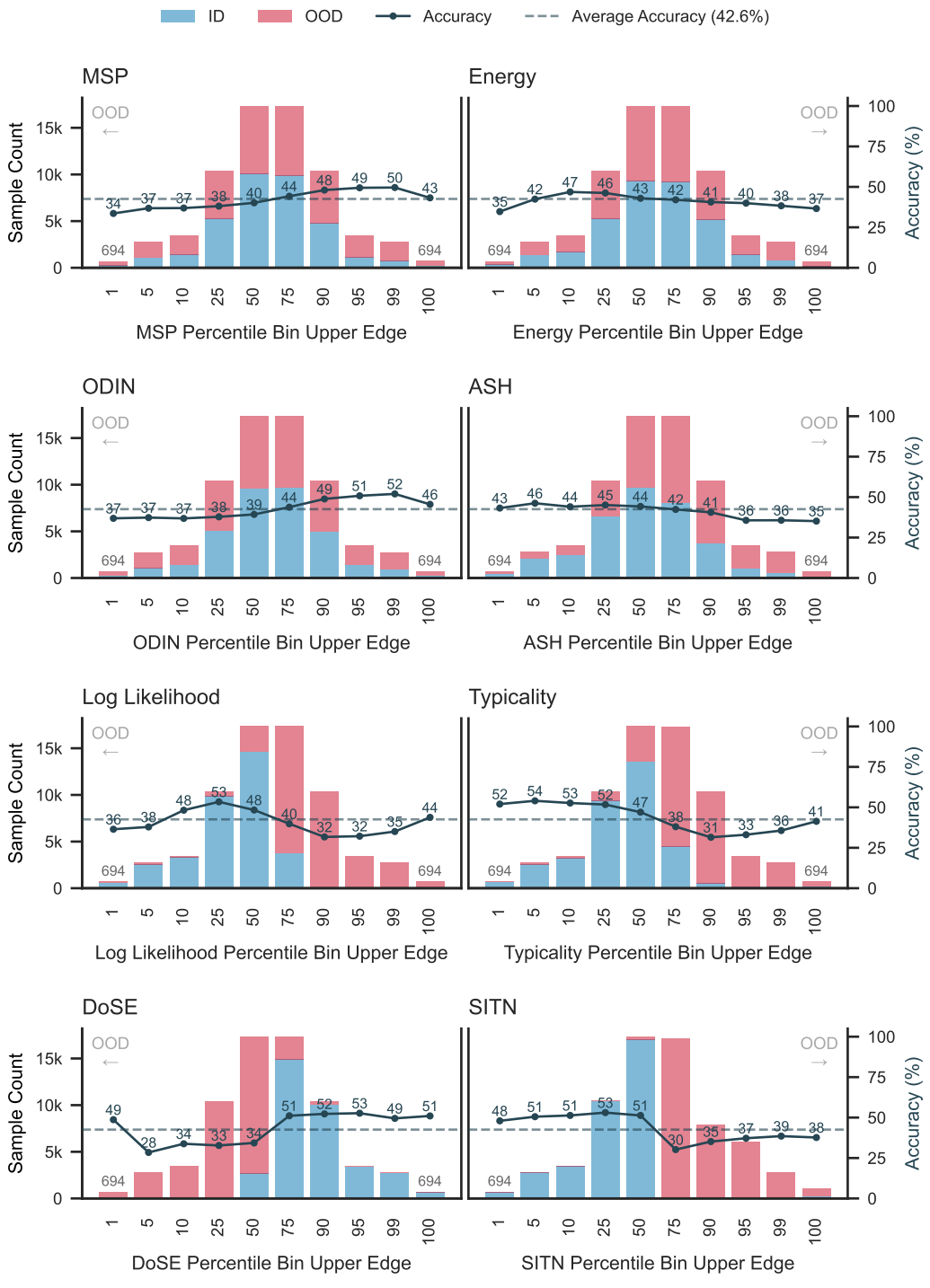

Figure 18: **Dementia diagnosis on CAUEEG — resampling perturbation (200 Hz)**

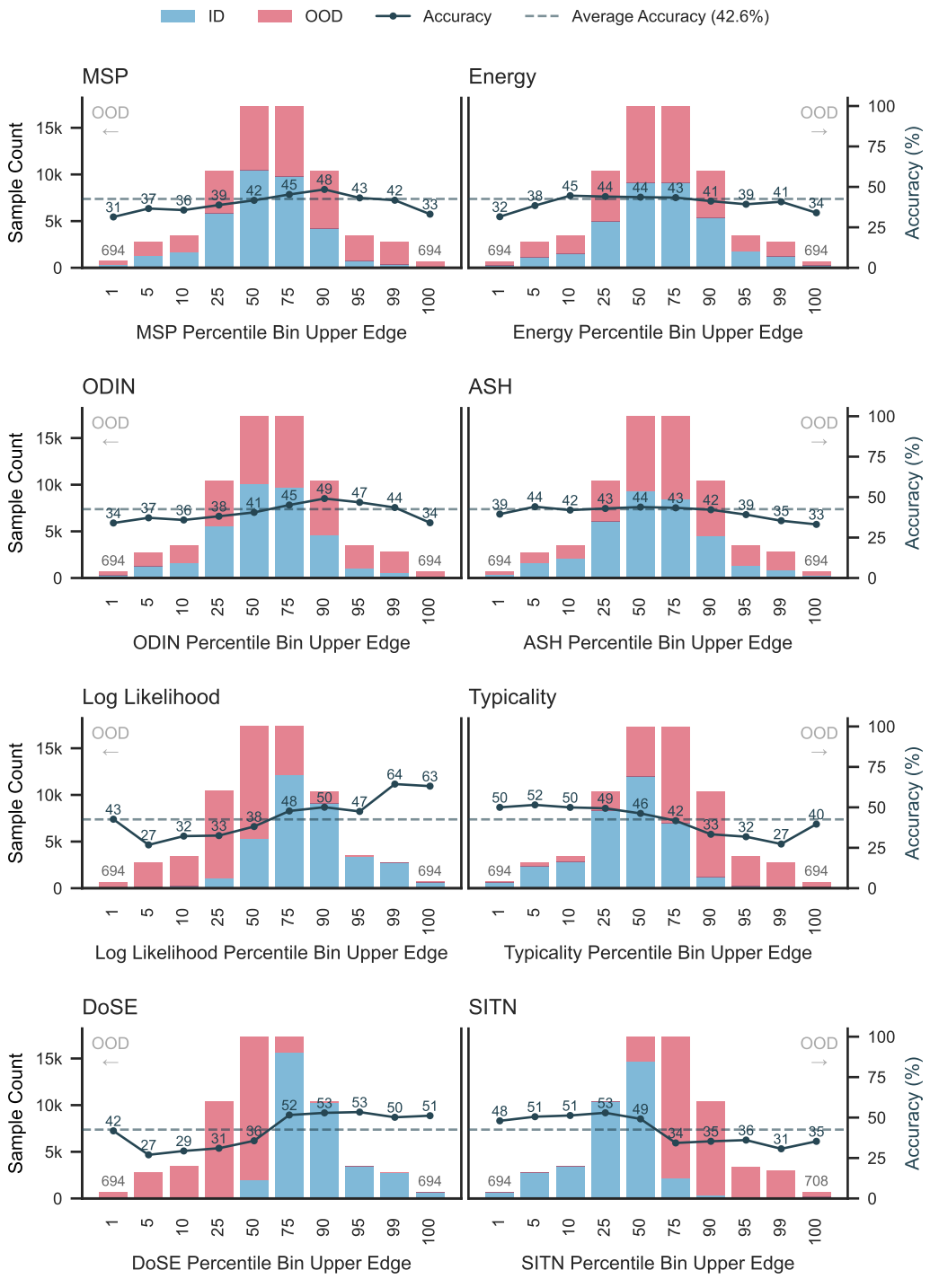

Figure 19: **Dementia diagnosis on CAUEEG — channel shuffle perturbation (100%)**

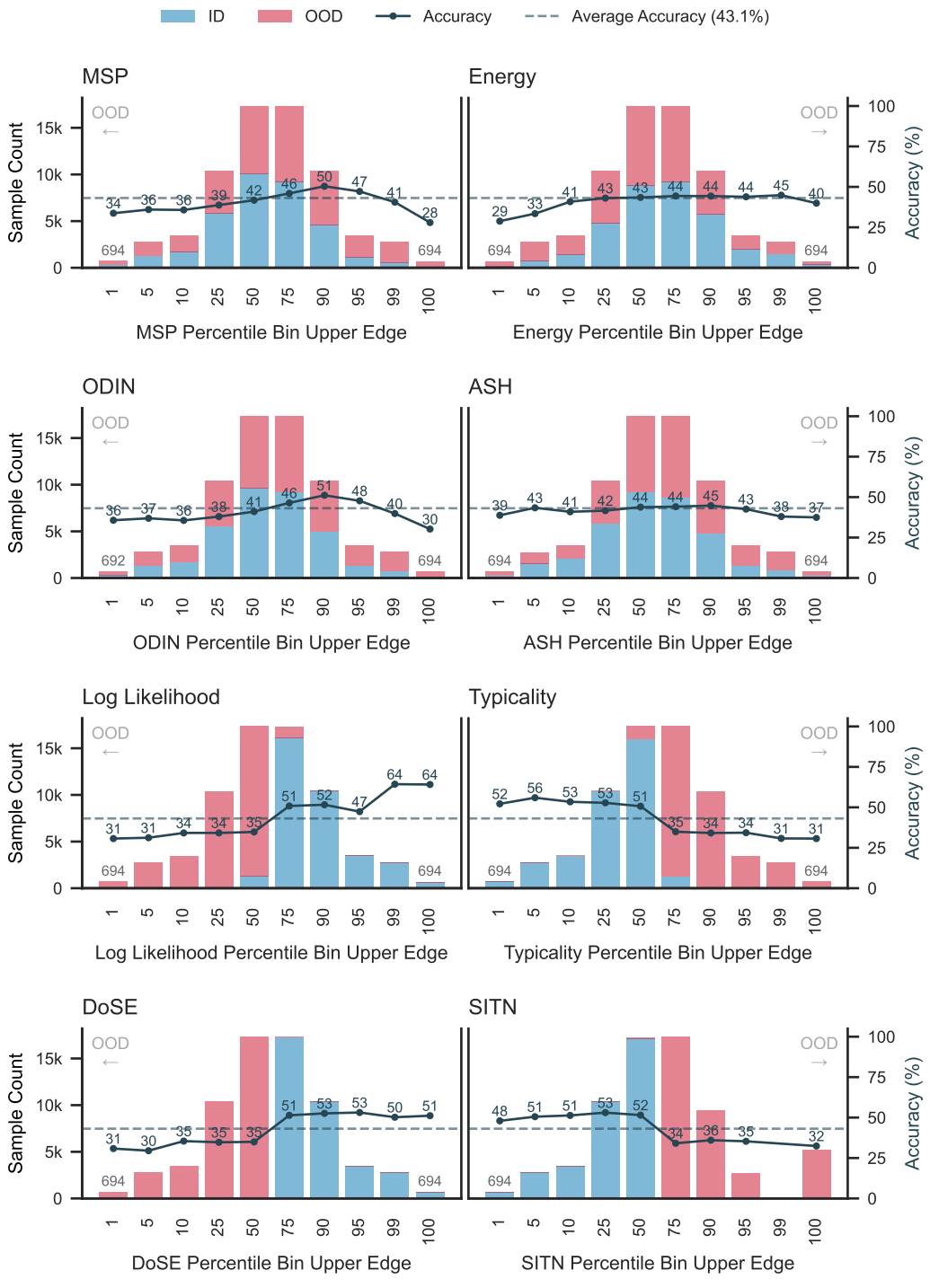

Figure 20: **Dementia diagnosis on CAUEEG — re-referencing perturbation (bipolar)**
Note: The irregular bin sizes at the upper percentiles of the SITN plot are due to metric saturation, where a large number of tied maximum scores are grouped into the final bin.

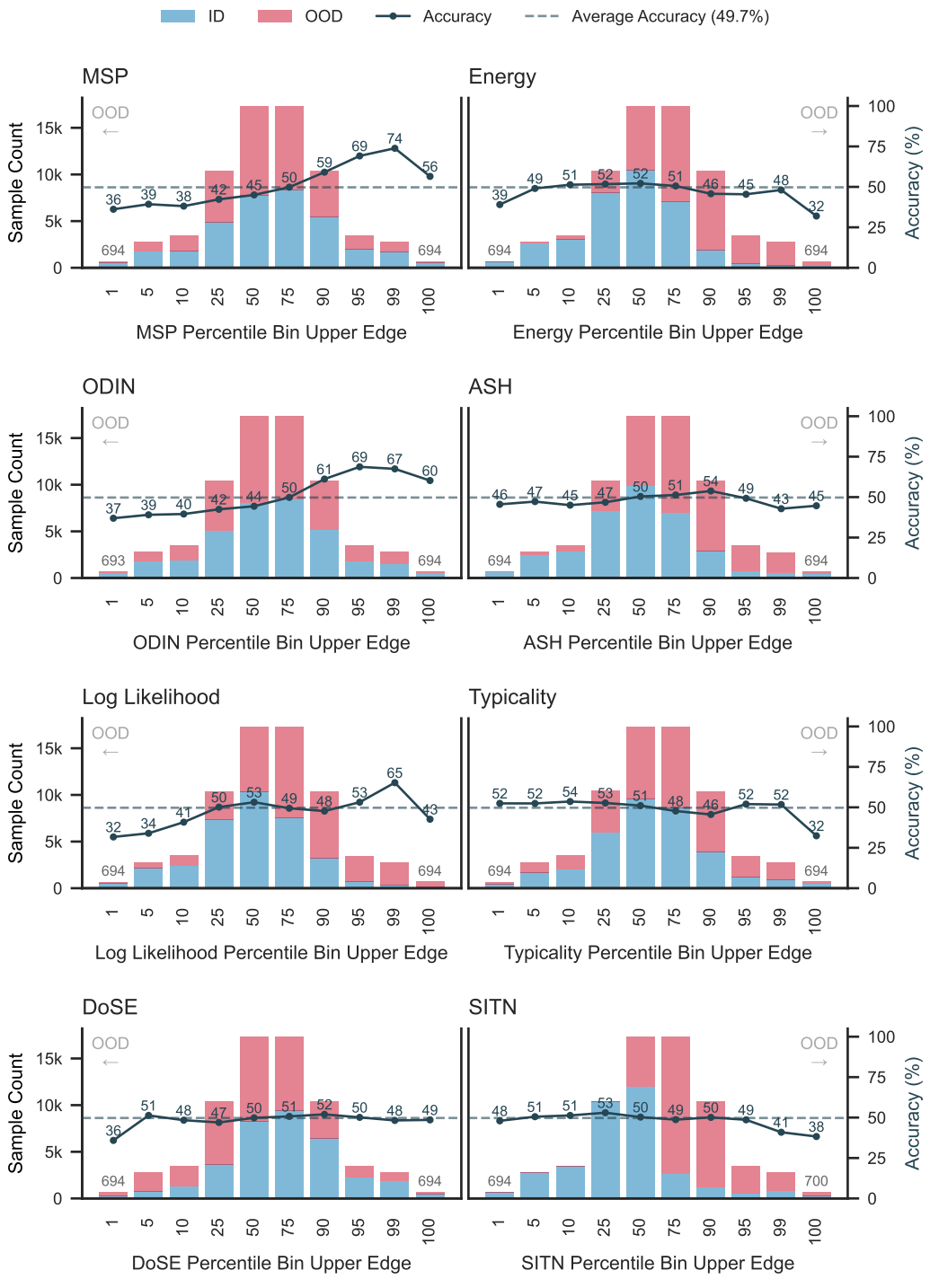

Figure 21: **Dementia diagnosis on CAUEEG — high-pass filter perturbation (8 Hz)**

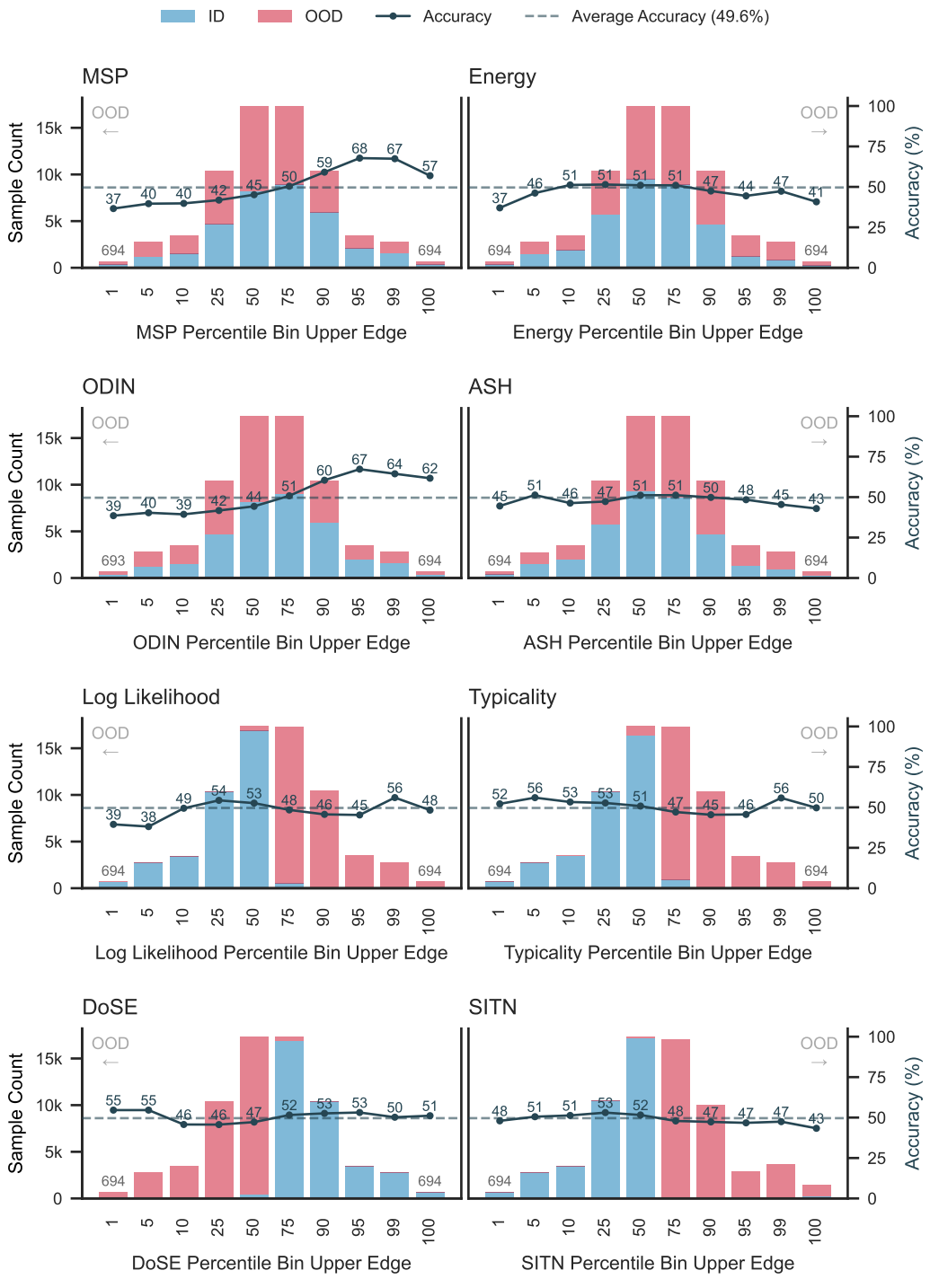

Figure 22: **Dementia diagnosis on CAUEEG — low-pass filter perturbation (15 Hz)**