

THE LANCET

Digital Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed.
We post it as supplied by the authors.

Supplement to: Sun C, Karakis I, Herlopian A, et al. Toward unified and comprehensive automated electroencephalogram interpretation: a multicentre development and validation of an electroencephalogram foundation model. *Lancet Digit Health* 2026.
<https://doi.org/10.1016/j.landig.2026.101039>

SUPPLEMENTAL MATERIAL for

Toward Unified and Comprehensive Automated EEG Interpretation

Contents

SUPPLEMENTAL MATERIAL FOR	1
1. DATASETS	2
1.1 Pre-training dataset	3
1.2 Fine-tuning datasets	6
1.3 Internal Holdout Validation datasets	7
1.4 External Validation datasets	8
2. DEVELOPMENT OF THE AI MODEL	12
2.1 EEG Input	15
2.2 EEG Tokenizer	16
2.3 EEG Transformer	17
2.4 Task Heads	18
3. BASELINES	19
4. STATISTICAL ANALYSES	20
5. RESULTS FOR SINGLE-SCORED VALIDATION DATASETS	23
6. RESULTS FOR MULTI-EXPERT VALIDATION DATASETS	27
7. INTER-RATER RELIABILITY	29
8. CALIBRATION ANALYSIS	31
9. CASE STUDIES	34
10. COMPREHENSIVE RESULTS	50
11. SUBGROUP RESULTS	71
11.1 Effects of on sex on algorithm performance	71
11.2 Age group	79
12. VALIDATION ON DIVERSE INTERNATIONAL COHORTS	121
13. TEST ON MISSING CHANNELS DATASETS	126
REFERENCES	129

Tables

- Table S1. EEG Interpretation Tasks
- Table S2. EEG Findings for Model Development and Validation
- Table S4. Comprehensive Results for 10 datasets

Figures

- Figure S1. Global distributions of scoring experts and datasets
- Figure S2. Characteristics of Model Development Datasets
- Figure S3 Model Architecture
- Figure S4. Performance Results on Single-scored Datasets.
- Figure S5. Additional Performance Results on Multi-expert Scored Datasets
- Figure S6. Additional Reliability Analyses
- Figure S7. Additional calibration curves
- Figure S8-S21. Case studies

1. Datasets

As listed in Table 1, we curated a large EEG dataset, the Harvard Electroencephalography Database (HEEDB)¹ for pre-training and fine-tuning the model. We utilized 7 external holdout datasets to validate model performance. The data is divided into three parts: pretraining, finetuning, and validation. A data partitioning consort diagram is shown in Figure 1. For model development and training we utilized 12 (HEEDB, IIC, SN2, MGH-PSG, BCH-PSG, MoE, TUH, MASS, UPenn, SAI, ON, HEP) datasets from a total of 34,602 patients collected from 42 different institutions in 8 countries and 3 continents (Figure S1A). Ground truth labels were contributed by 110 EEG experts from 62 different institutions in 11 countries and 4 continents (Figure S1B).

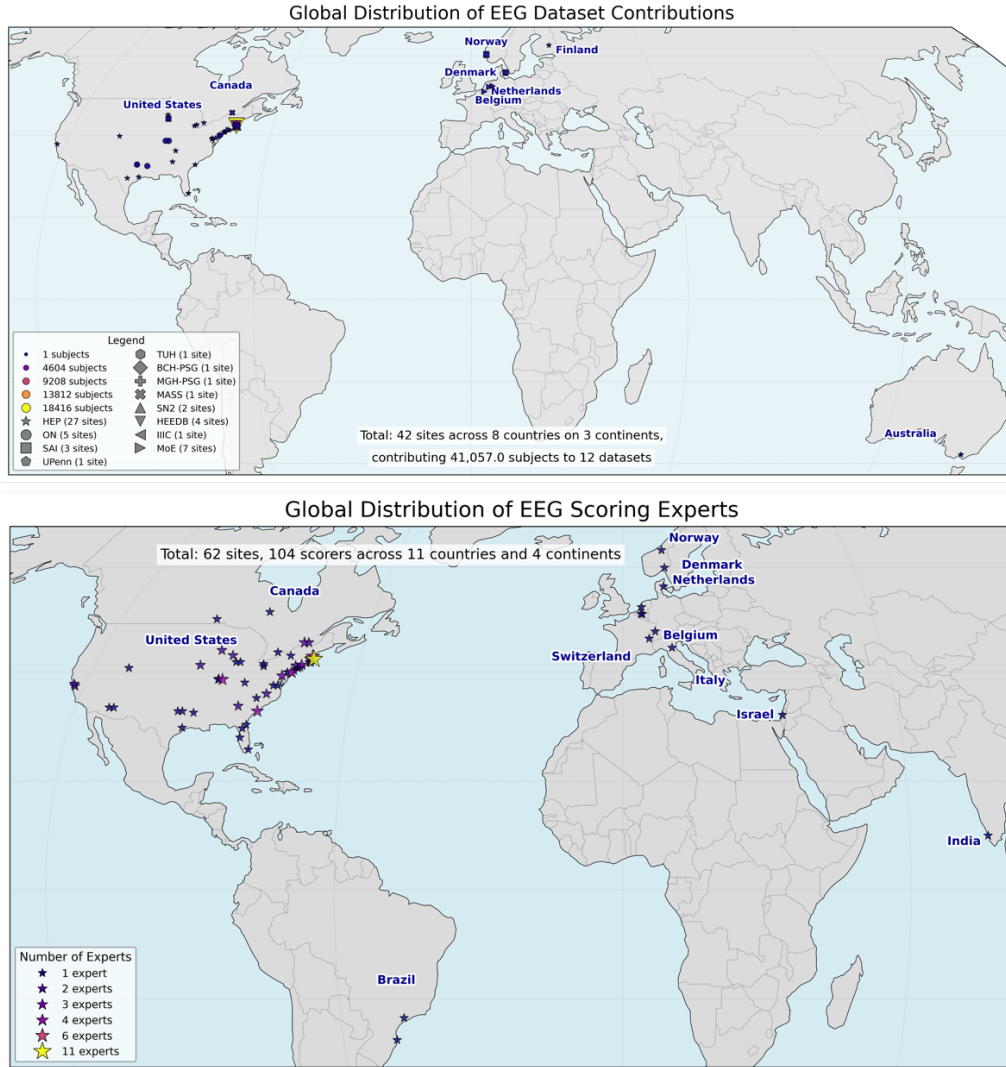


Figure S1. Global distributions of (A) datasets and (B) scoring experts.

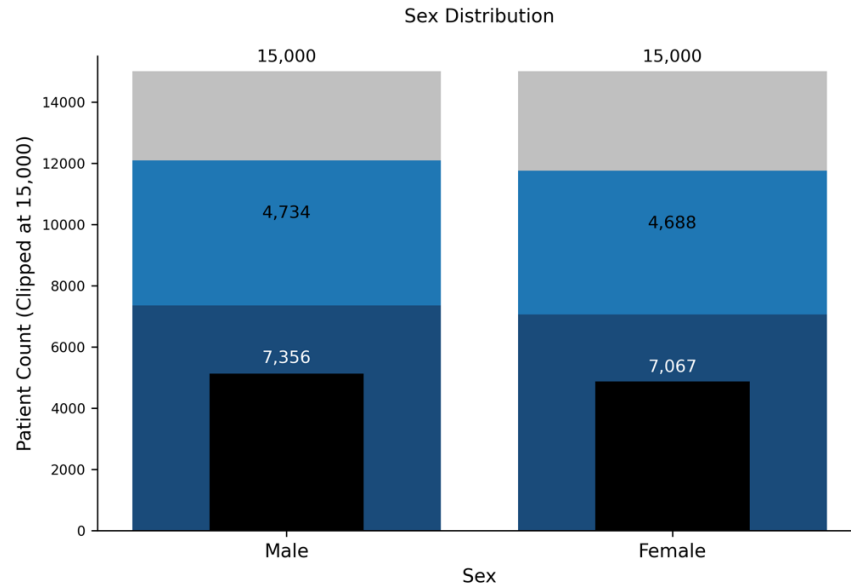
1.1 Pre-training dataset

The Harvard EEG Database (HEEDB): is a large-scale, cloud-based, research resource currently comprising 284,341 EEGs from 109,177 patients across four Harvard-affiliated hospitals – Massachusetts General Hospital (MGH), Brigham and Women’s Hospital (BWH), Beth Israel Deaconess Medical Center (BIDMC), and Boston Children’s Hospital (BCH). EEGs in HEEDB include routine outpatient EEGs (rEEG) for diagnostic test and epilepsy management, continuous EEGs recorded in the acute care and ICU settings, and EEGs recorded in Epilepsy Monitoring Units (EMUs).

From HEEDB, we selected one EEG from each of 15,000 unique patients across rEEG, ICU, and EMU settings, and extracted 10-minute EEG data per patient, resulting in a dataset of approximately 2,400 hours for model pre-training.

We selected this subset following principles derived from our scaling law analysis²⁵. According to scaling laws, model size (number of parameters) and the required training dataset size generally follow a sublinear relationship. Specifically, to maintain training efficiency (i.e., achieve optimal performance under a fixed compute budget), the model size N and dataset size D roughly satisfy: $D \propto N^{0.73}$. Given the model size of 41 million parameters ($15.04 \propto 41^{0.73}$), we aimed to balance dataset size with computational efficiency and generalization capacity. Empirically, we found that using approximately 10-minute EEG segments per patient — totaling around 14,000 patients — provided an optimal volume of data for pre-training at this scale. This approach allowed us to maximize the diversity across patients while avoiding excessive overfitting risks or unnecessary computational burden. Altogether, this subset comprised roughly 2,333 hours of EEG data, translating to approximately 16 million tokens, where each 1-second single-channel segment was treated as an individual token. This strategy ensured a robust representation of EEG patterns across a wide range of clinical scenarios, which is critical for developing a generalizable foundation model.

To maximize diversity in EEG findings, we pre-selected data using EEG reports, clinician annotations, and other algorithmic methods, followed by expert validation. As illustrated in Figure S2, the pretraining dataset encompasses a diverse range of sex, ages and EEG findings.



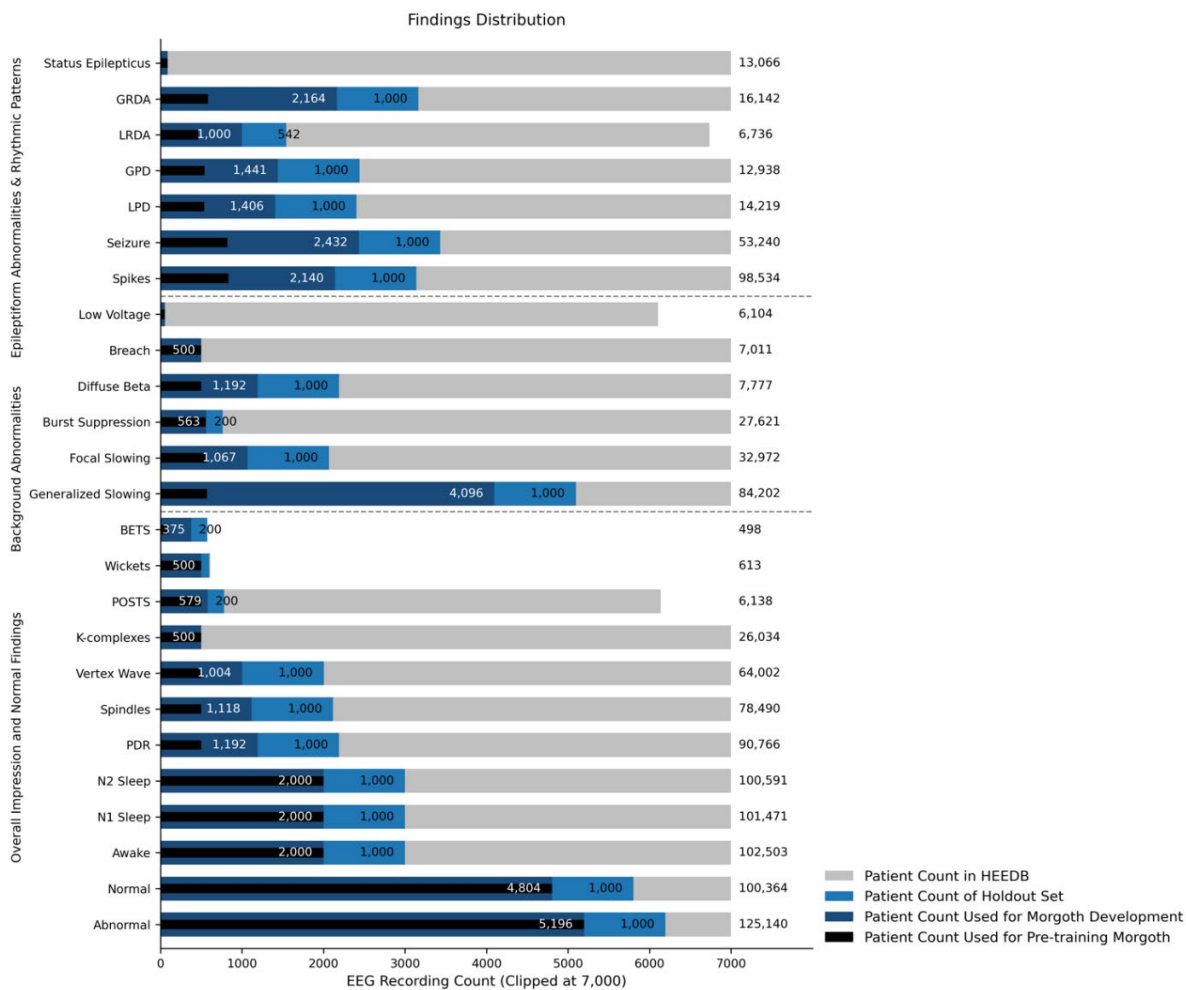
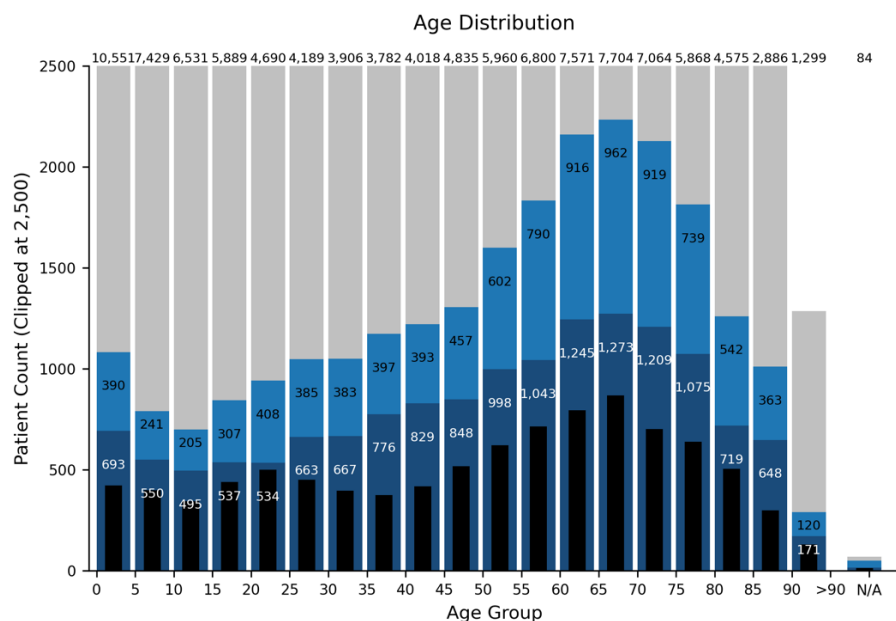


Figure S2. Characteristics of Model Development Datasets.

1.2 Fine-tuning datasets

Fine-tuning also relied on data from HEEDB, and fine-tuning MORGOTH for downstream tasks required a large volume of labeled training data. However, the EEG-related labels available in hospital-acquired datasets are limited to EEG reports and clinician annotations. EEG reports provide a global summary of the EEG without event-level localization, while time-stamped annotations from clinicians within EEGs typically lack authoritative validation. Therefore, a systematic labeling process is essential to construct a well-annotated dataset that enables robust supervised learning.

As a start, we carefully prepared the fine-tuning dataset for MORGOTH from scratch. Our fine-tuning dataset includes labels derived from EEG reports or clinician annotations, as well as EEG events annotated by human experts. To minimize the effort required from experts, we developed a structured labeling workflow that ranged from EEG report and annotation preparation to algorithmic processing and, finally, expert validation. This process resulted in the creation of the internal dataset, as summarized in Figure 1. Additionally, for most tasks, the fine-tuning data included at least 2,000 patients, with 1,000 allocated for training and 1,000 for testing.

1.2.1 Fine-tuning data labeled via clinical care plus one additional expert

For the overall interpretation of EEG findings, we extracted information from the impressions in the EEG reports using regular expressions to assign a label to each EEG as either “normal” or “abnormal”.

For sleep staging (awake, N1, N2, N3, REM), we defined two tasks. The first task focuses on distinguishing between the awake, N1, and N2 stages. Since most clinical EEGs lack dedicated annotations for all five stages, labels for these three stages were inferred from EEG annotations such as “awake,” “drowsy” (indicating N1), and “sleep spindles” (indicating N2). The second task classifies all five sleep stages, using expert-labeled polysomnography (PSG) data from MGH², where each recording was labeled by one of seven sleep technicians and later edited by a board-certified sleep physician. However, the PSG data is limited to six channels from the 10-20 system (F3, F4, C3, C4, O1, and O2). As a result, only these six EEG channels were used during fine-tuning for this task.

For background abnormalities, we utilized Medical LLaMA³ to screen EEG reports applying a structured prompt: “*Based on the medical context of neurology and EEG analysis, does the following report indicate the presence of [finding] in the EEG? Respond with yes, no, or not mentioned.*” – this process provides an initial set of EEG-level labels. We then refined the timing of the events of interest using time-stamped free-text EEG annotations from clinicians. Furthermore, we utilized a previously validated automated

burst suppression detector⁴ to identify segments that were highly likely to contain burst suppression patterns. Finally, the extracted data were processed through our annotation platform, where an expert (MBW) visually reviewed and assigned a final label to the EEG samples, establishing the final ground truth.

For epileptiform abnormalities and rhythmic patterns, we followed the same procedure as described above. After the initial screening of EEG reports and annotations, SpikeNet⁵ and SPaRCNet⁶ were used to identify 1-second spike events and 10-second segments of seizure, lateralized periodic discharges (LPD), generalized periodic discharges (GPD), lateralized rhythmic delta activity (LRDA), and generalized rhythmic delta activity (GRDA), which were then validated by one expert (MBW).

1.2.2 Fine-tuning data labeled by multiple experts

In addition, we incorporated the IIIC and SN2 datasets, both sourced from the HEEDB database, as part of our internal datasets. These datasets contain labels from multiple experts in prior studies^{7,8} and provide soft labels that approximate class probabilities. To clarify subsequent testing tasks and distinguish between annotating experts, we divided the HEEDB at the patient level into HEEDB-Train/Test, IIIC-Train/Test, and SN2-Train/Test sets.

The IIIC dataset⁷ comprises 47,330 10-second EEG segments from 2,698 patients at MGH (as part of the HEEDB dataset). These events were annotated by 124 experts, with over 80% of seizure, LPD, GPD, LRDA, and GRDA labels (21,760 out of 26,365) assigned by 2 to 23 experts. For model training, we extracted 28,465 segments from 1,940 patients (49% female, median age 54 years, range 0 to 96), creating the IIIC-Train dataset.

The SN2 dataset⁸, a subset of HEEDB, comprises 64,476 one-second EEG segments from 5,049 patients at MGH and BWH. It is called “SN2” because it was used in the development and testing of the model “SpikeNet 2.0”. The SN2 dataset includes 11,641 spike events annotated by at least 8 experts from a group of 24 experts, as well as 52,862 control/negative events labeled by at least one expert. For model training, we extracted 10,651 segments from 4,886 patients (49% female, median age 53 years, range 0 to 96), creating the SN2-Train dataset.

1.3 Internal Holdout Validation datasets

To validate MORGOTH, we ensured that the internal holdout validation datasets had no patient-level overlap with the training and fine-tuning datasets. Specifically, the internal holdout validation dataset includes the HEEDB-Test, IIIC-Test, SN2-Test, and MGH-PSG-Test datasets. Additionally, we incorporated the BCH-PSG dataset and data from MGH, BWH, and BIDMC

within the MoE dataset. Among these, the IIIC-Test, SN2-Test, and MoE datasets feature multi-expert scored labels.

The Map of Everything (MoE) dataset is a multi-expert annotated dataset designed for comprehensive event-level evaluation. It includes 2,761 EEGs from 2,250 patients across MGH, BWH, and BIDMC (considered part of the internal validation dataset), as well as data from the International Cardiac Arrest Research Consortium Electroencephalography (I-CARE)^{9,10}, which covers ULB, UTW, and YNG (external datasets, but labeled in conjunction with the rest of MoE). MoE spans 46 distinct EEG patterns, with 21 experts contributing to the labeling process. Each event was verified by an average of 13 experts, with a range of 10 to 19 experts per event.

1.4 External Validation datasets

We also incorporated seven additional external validation datasets, which allow for the evaluation of one or more event-level or EEG-level tasks.

The Human Epilepsy Project (HEP) dataset is a 6-year, 29-site study of 18 to 60-year-old patients with newly diagnosed focal epilepsy, containing 143 EEGs. Each EEG was independently labeled by 3 out of 14 experts, who marked all detected spikes. Events with at least 2 out of 3 “yes” votes were considered positive, while those with only 1 “yes” vote were considered ambiguous and omitted. Since the three experts assigned were not the same across all EEGs, it was not possible to plot individual expert points or calculate EUC performance.

The Occasion Noise dataset is a multi-label dataset containing 100 EEGs from 100 patients, with EEG-level annotations from 18 experts. The data is sourced from Barnes-Jewish Hospital (42 patients), Washington University in St. Louis (21 patients), Dallas Veterans Affairs Medical Center (15 patients), University of Texas Southwestern Medical Center (14 patients), and Louisiana State University (8 patients). The dataset includes classifications for focal and generalized non-epileptiform abnormalities (such as focal and generalized slowing), as well as focal and generalized epileptiform discharges.

The SAI dataset¹¹ is the holdout set of the SCORE-AI model and includes 100 patients from Haukeland University Hospital, the Danish Epilepsy Center, and Mayo Clinic. It features EEG-level labels from 14 experts, including classifications for normal vs. abnormal, epileptiform-focal (focal spikes), epileptiform-generalized (generalized spikes), nonepileptiform-focal (focal slowing), and nonepileptiform-generalized (generalized slowing). Event-level labels are not available in SAI dataset.

The Temple University EEG Corpus (TUH)¹² is a widely used open-source dataset for model evaluation. In this study, we focus on the TUAB, TUEV, TUSL, and TUEP tasks. Since the task definitions in TUH differ from ours, we retrained a MORGOTH head on the TUH-designated training set to ensure a fair comparison with other models evaluated on TUH. This approach enables us to assess the capabilities of the pre-trained base model and compare its performance against existing EEG foundation models on TUH benchmarking tasks.

The UPenn Dataset¹³ comprises 69 PSG recordings from 69 individuals, conducted in clinical sleep lab settings. Each PSG was independently sleep-staged by 6 sleep technicians from sleep labs at the University of Pennsylvania, St. Luke’s Hospital, and Stanford University. In this study, we utilized only the 2 EEG channels.

The Montreal Archive of Sleep Studies (MASS) dataset¹⁴ is an open-access, collaborative database containing laboratory-based PSG recordings from three sleep laboratories. This cohort includes PSG data from 200 full-night recordings of 97 men and 103 women, aged 18 to 76 years (mean: 38.3 years, SD: 18.9 years). In this study, we utilized only the 16 EEG channels.

Table S1. EEG Interpretation Tasks.

Interpretation	EEG findings	Event-level task	Event-level Evaluation	EEG-level task	EEG -level Evaluation
Overall Interpretation					
Normal vs. abnormal	Normal, abnormal	Binary classification	Binary	1 binary classification	Binary
Sleep staging	Awake, N1, N2, N3, REM	5-class classification	One-vs-rest	5 binary classifications	Binary
Background					
Slowing	Focal slowing, generalized slowing	3-class classification	One-vs-rest	2 binary classifications	Binary
Burst suppression	Burst suppression	Binary classification	Binary	1 binary classification	Binary
Epileptiform Abnormalities & Rhythmic Patterns					
Spikes detection	10-second spike	Binary classification	Binary	1 binary classification	Binary
Spike localization	Focal spike, generalized spike	3-class classification	One-vs-rest	2 binary classifications	Binary
Seizure and IIC	Seizure, LPD, GPD, LRDA, GRDA	6-class classification	One-vs-rest	5 binary classifications	Binary

Table S2. EEG Findings for Model Development and Validation.

Internal Datasets			External Datasets	
HEEDB ¹ (Event-level; 1-3 expert)			MoE (Event-level; 21 experts)	
Class	Train	Test	Class	Evaluation

Total pre-training	14,500	---	Total	2,250(External 409)
Total fine-tuning	18,677	14,352	Burst suppression	179(42)
Normal	4,000	1,000	Focal slowing	200(2)
Abnormal	4,000	1,000	Generalized slowing	1,029(202)
Burst suppression	563	200	Focal spikes	102(2)
Focal slowing	1,067	1,000	Generalized spikes	72(29)
Generalized slowing	4,096	1,000	Spikes	174(31)
Spikes	694	959	Seizure	23(1)
Focal spikes	1,662	1,000	LPD	68(2)
Generalized spikes	2,057	1,000	GPD	54(28)
Seizure	1,507	653	LRDA	17(0)
LPD	548	606	GRDA	14(5)
GPD	1,061	642	Awake	301(3)
LRDA	560	186	Sleep N1	32(0)
GRDA	1,146	536	Sleep N2	106(1)
Awake	1,000	1,000	Sleep N3	0(0)
Sleep N1	1,000	1,000	REM	5(0)
Sleep N2	1,000	1,000	SAI (EEG-level; 14 experts)	
BETS	375	200	Total	100
POSTS	579	200	Normal	68
Spindles	1,118	1,000	Focal slowing	21
Wickets	111	100	Generalized slowing	34
Vertex wave	1,004	1,000	Focal spikes	26
PDR	1,192	1,000	Generalized spikes	20
SN2 ² (1-second Event-level; 24 experts)			Occasion Noise (Event-level; 15 experts)	
Total	4,886	208	Total	100
Spikes	1,466	41	Focal slowing	12
Spike-free	3,420	167	Generalized slowing	18
Spike segment	7,231	1,053	Focal spikes	15
None-spike segment	54,774	1,418	Generalized spikes	16
IIC ² (Event-level; 30 experts)			HEP (1-second event-level; 14 experts)	
Total	1,819	743	Total	143
Seizure	925	347	Spikes	51
LPD	498	394	Spike-free	92
GPD	425	358	Spikes segment	1,650
LRDA	440	256	None-spike segment	1,646

GRDA	1,018	464	TUH ³ test set (Event-level; 1 expert)	
Others	1,218	608	Total	552
Seizure segment	11,647	5,283	Normal	150
LPD segment	4,138	3,950	Slowing	33
GPD segment	2,081	854	Spikes	36
LRDA segment	762	529	Seizure	43
GRDA segment	1,158	896	LPD	81
Others segment	8,679	7,353	GPD	79
MGH PSG (Event-level; 1 expert)			UPenn (Event-level; 6 expert)	
W, N1, N2, N3, REM	1,000	1,000	W, N1, N2, N3, REM	69
BCH PSG (Event-level; 1 expert)			MASS (Event-level; 1 expert)	
W, N1, N2, N3, REM	/	1,000	W, N1, N2, N3, REM	200

The counts in table are based on the number of patients. The patients in the training, testing, and evaluation sets are strictly non-overlapping. The model is not trained on external datasets but is directly evaluated on them. In the internal dataset, the class labels are annotated at the event level, with 10-segment events and 1-second spike events. ¹HEEDB dataset are used for pre-training and fine-tuning, the task-related features are used as the positive set while other features are used to construct the control set. ²SN2 dataset and IIIC dataset are derived from HEEDB. They are well-established datasets from SpikeNet and SPaRCNet. Therefore, we present them separately for clarity and distinction. ³TUH evaluation includes 3 tasks: TUAB(Normal/Abnormal), TUSL(Seizure/Slowing/Background), and TUEV(Spikes/GPD/LPD/Eye movement/Artifact/Background). For comparison with other models, we need to train on the train set provided by the dataset and test on the test set.

Overall, for internal validation, we utilized 13,618 EEGs from 13,334 patients (51% male, 49% female; median age 59 years with range: 0 to 90+ years) from four hospitals, annotated by 1-30 experts. For external validation, we used seven datasets comprising 1,800 EEGs from 1,573 patients (51% male, 49% female; median age 45 years with range: 0 to 90+ years) from 48 medical institutions. Our test sets were divided into separate subsets. While different subsets might include EEGs from the same patient, within each individual subset, only a single EEG per patient was used for evaluation to ensure the independence of samples within that subset. For example, as shown in Table 1, for the internal test sets (HEEDB-Test, IIIC-Test, SN2-Test, MGH-PSG-Test, and BCH-PSG), the number of patients equals the number of EEGs, confirming that each patient contributes only one EEG to the evaluation within each subset. The only exception is the MoE-Internal dataset, where individual patients may have more than one EEG. However, this dataset was designed to evaluate multiple distinct tasks, and for each task, we ensured that the evaluation was performed at the patient level.

263

264 2. Development of the AI Model

265 We develop the model to achieve 7 event-level tasks and 17 EEG-level tasks, as listed in Table S1.
266 The overall model structure is shown in Figure S3. It is an EEG foundation model, having a
267 tokenizer, an EEG Transformer, and multiple task heads. First, the tokenizer is trained using
268 contrastive learning and vector-quantized prediction. It converts continuous EEG signals into
269 discrete tokens and builds a code book (a set of unique vectors) for EEG tokens. Then, the EEG
270 Transformer is pre-trained through masked token prediction, allowing the model to capture both
271 temporal and spatial EEG features and to learn relationships among the tokens. Finally, we fine-
272 tune the model with added task heads on specific tasks using supervised learning strategies. The
273 event-level heads are directly connected to the original model for training, whereas EEG-level
274 heads are built and trained on the continuous outputs of the event-level predictions. The model
275 hyperparameters are shown in the Table S3.

276

277

278

279

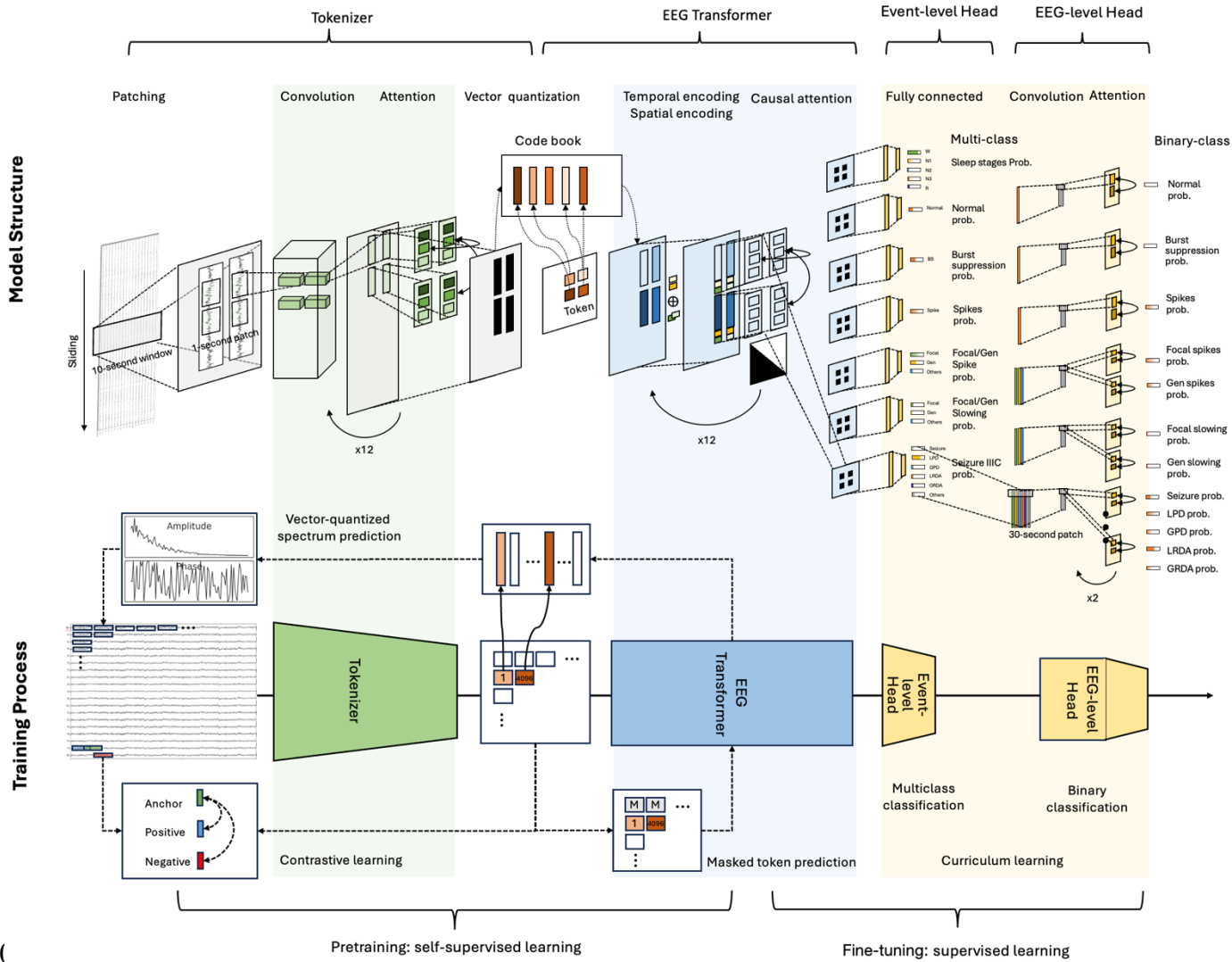


Figure S3: Model Architecture

Table S3. Model Hyperparameters

Model Structure	Hyperparameters	Values
Tokenizer and VQ Pre-training		
Convolution Encoder	Input channels	{1,8,8}
	Output channels	{8,8,8}
	Kernal size	{15,3,3}
	Stride	{8,1,1}
	Padding	{7,1,1}
Transformer	Encoder layers	12
	Decoder layers	3
	Hidden size	200
	MLP size	800

	Attention head number	10
Codebook	Size	8192×64
Pre-training	Batch size	512
	Learning rate	[1e-5,5e-5]
	Learning rate scheduler	Cosine
	Optimizer	AdamW $\beta(0.9,0.99)$
	Weight decay	5e-5
	Total/Warmup epoch	120/20
Transformer and masked EEG Pre-training		
Transformer	Encoder layers	12
	Hidden size	200
	MLP size	800
	Attention head number	10
Pre-training	Batch size	512
	Learning rate	[1e-5,5e-4]
	Learning rate scheduler	Cosine
	Optimizer	AdamW $\beta(0.9,0.98)$
	Weight decay	5e-3
	Total/Warmup epoch	100/10
	Data stride	500
	Gradient clipping	3
	Layer scale init	0.1
	EMA weight	0.996
	Mask ratio	0.5
One Event-level Task Head (A total of 7) and Downstream Fine-tuning		
Transformer	Encoder layers	6
	Hidden size	200
	MLP size	800
	Attention head number	10
Fully connected layer	Hidden size	{256, 128, 56}
	Layer number	3
Fine-tuning	Batch size	64
	Learning rate	[1e-6,1e-4]
	Learning rate scheduler	Cosine
	Optimizer	AdamW $\beta(0.9,1)$
	Weight decay	5e-2
	Total/Warmup epoch	20/3
	Drop path	0.1
	Layer-wise learning rate decay	0.65
One EEG-level Task Head (A total of 17) and Downstream Fine-tuning		
Convolution Encoder	Layers	2
	Input channels	{event_class_number*, 16}
	Output channels	{16,32}

	Kernal size	{3,3}
	Padding	{10,3}
Positional Encoding	Channels	15000 × 32
Transformer	Encoder layers	2
	Hidden size	64
	MLP size	{32,64}
	Attention head number	4
Fine-tuning	Batch size	64
	Learning rate	[1e-6, 1e-4]
	Learning rate scheduler	Cosine
	Optimizer	AdamW $\beta(0.9, 1)$
	Weight decay	5e-2
	Total/Warmup epoch	20/3
	Dropout	0.1

* event_class_number indicates the number of classes for the corresponding event-level task.

2.1 EEG Input

EEG signals can be represented as a matrix $X \in \mathbb{R}^{C \times T}$ where C denotes the number of EEG channels and T represents the total number of signal points, determined by the recording duration and sampling rate.

Most EEGs in our data included 19 channels ($C = 19$) based on the international 10-20 system: Fp1, Fp2, F3, F4, C3, C4, P3, P4, F7, F8, T3, T4, T5, T6, O1, O2, Fz, Cz, Pz. Exceptions were EEGs recorded as part of PSG, which contained a subset of just 6 channels (F3, F4, C3, C4, O1, O2). All signals were resampled to 200 Hz, and band-pass filtered between 0.5 Hz and 70 Hz. Additionally, notch filtering was applied at 50 Hz and 60 Hz to remove the power-line interference. EEG signals are in microvolts (μV) and converted to common average montage. The signals were clipped at $\pm 500 \mu V$ and normalized to a $[-1, 1]$ range before being input into the model to reduce artifacts.

Generally, each EEG can be segmented into $L = \left\lfloor \frac{T-w}{s} \right\rfloor + 1$ segments if the window-size is w and the sliding-step is s . In our approach, $w = 10s \times 200Hz$, an EEG X is first chunked into 10-second 19-channel segments $\{X_i\}_{i=1}^L$, then a segment X_i is patched into 10 non-overlapping 1-second 1-channel patches $\{x_i\}_{i=1}^{10 \times 19}$. Before feeding these patches into the Transformer module for patch-wise interaction via self-attention, a temporal encoder¹⁵, consisting of 1-D convolution kernels, GELU activation, and a normalization layer, was employed to encode patches into embeddings $\{h_i\}_{i=1}^{10 \times 19}$. This encoding process captures temporal features, addressing the high temporal resolution of EEG signals.

2.2 EEG Tokenizer

The tokenizer in language models splits the input text into the smallest units that the model can process (e.g., words, sub-words, or characters) and maps them to corresponding token IDs^{16,17}. Inspired by this, several studies have applied a similar approach in Transformer-based EEG models^{59,60}. The EEG tokenizer is trained prior to the pretraining of the EEG Transformer using a self-supervised learning strategy based on Vector Quantization (VQ)¹⁸⁻²⁰. The codebook contains all tokens, where the VQ maps each patch representation to its nearest neighbor. In our approach, the tokenizer consists of 12 Transformer encoder blocks with 10 attention head, and each EEG token represents a patch e . After processing through the tokenizer, it is mapped to a vector v_i retrieved from the codebook $V = \{v_i \mid i = 1, \dots, K\} \in R^{K \times D}$, where K is the number of discrete vectors, and D represents the dimension. In Morgoth setting, $K = 8192, D = 128$. For the i th token, the process of obtaining its quantized code can be formally expressed as $z_i = \arg \min_j \|e_i - v_j\|_2$, where $e_i = f(h_i)$ and f represents the function of tokenizer's network.

To train the VQ-based tokenizer, we use the Fourier spectrum prediction method, which enhances training convergence, as direct reconstruction of raw EEG signals fails to converge in most experiments¹⁸. The spectrum prediction method based on VQ trains the tokenizer network by minimizing the discrepancy between the predicted amplitude $\hat{A}$ and phase $\hat{\phi}$ regressed through a regression layer after 3 Transformer decoders, where the ground-truth amplitude A and phase ϕ are obtained via the discrete Fourier transform²¹. However, modeling only the amplitude and phase may lead to the loss of essential EEG signal information. For example, it may overlook transient features like spikes, slow waves, and rhythmic patterns, potentially causing spikes and some benign spike-like waves to share the same quantized code. This issue also makes it challenging to effectively activate all codes in a large codebook during training. Therefore, we incorporated contrastive learning to differentiate waveform shapes²², using a loss that encourages similar samples x, x^+ to have closer representations e, e^+ in the embedding space while pushing dissimilar samples apart. We define a token x as the anchor instance. Its positives x^+ are the augmented instances obtained by the jitter-and-scale and permutation-and-jitter strategies²³. Its negatives x^- are from non-overlapping instances which do not have the same subsequence. Finally, given an EEG dataset $\mathcal{D} = \{X_i\}_{i=1}^N$, the VQ-based training loss, incorporating spectrum prediction and contrastive learning, is formulated as:

$$\mathcal{L}_{\mathcal{T}} = \sum_{x \in \mathcal{D}} \sum_i^L \left(\|\hat{A}_i - A_i\|_2^2 + \|\hat{\phi}_i - \phi_i\|_2^2 - \log \frac{\exp(\sigma(e_i, e_i^+))}{\exp(\sigma(e_i, e_i^+)) + \sum_{j=1}^B \exp(\sigma(e_i, e_{i,j}^-))} \right) + \|\mathbf{sg}[e_i] - v_{z_i}\|_2^2 + \|h_i - \mathbf{sg}[v_{z_i}]\|_2^2$$

where $\sigma(e, e^{+/-}) = \frac{\text{sim}(e, e^{+/-})}{\tau}$ is used to calculate the similarity between two vectors through a cosine similarity function sim with the temperature parameter τ , and $\mathbf{sg}$ represents the stop-gradient operation²⁰.

2.3 EEG Transformer

The tokenizer captures the signal characteristics of EEG patches but lacks their positional information within the entire EEG sequence. This includes both temporal information in the longitudinal time series and spatial information across channels. Therefore, in addition to the classical Transformer's temporal encoding $TE = \{te_i\}_{i=1}^{10}$, we introduce a spatial encoding list $SE = \{se_i\}_{i=1}^{19}$ to represent channel-specific information²⁴. The patch embedding is transformed into a concatenated representation $\text{concat}(e_i, te_i, se_i)$.

Subsequently, the sequence of embeddings is fed into the EEG Transformer having 12 encoder blocks with 10 attention heads. We incorporate more encoder structures rather than decoder structures, as encoders have demonstrated greater effectiveness for time series data in our experiments and align with the time series scaling law²⁵. The self-attention mechanism^{18,26} is defined as $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{LN(Q)LN(K)^T}{\sqrt{d_{head}}}\right)V$, where d_{head} represents the dimension of each attention head in the multi-head attention mechanism, and LN denotes layer normalization.

Pre-training is crucial as it enables the model to learn rich representations from large-scale data, improving generalization and downstream task performance. We employed self-supervised masked modeling strategies²⁷ for pre-training the EEG foundation model. The masks $\{m_i\}_{i=1}^N, m_i \in \{0,1\}$ were randomly generated for patch embeddings $\{e_i\}_{i=1}^N$. If $m_i = 1$, the corresponding embedding e_i is replaced as a learnable vector $e_i^{\mathcal{M}} \in R^D$. The masked embeddings $e^{\mathcal{M}} = \{e_i^{\mathcal{M}}\}_{i=1}^M \cup \{e_i\}_{i=1}^{N-M}$ are then processed through a Transformer encoder, which reconstructs the missing patches by predicting their original representations. The model is trained to minimize the reconstruction loss, defined as:

$$\mathcal{L}_{\mathcal{M}} = \frac{1}{\sum_{i=1}^N m_i} \sum_{i=1}^M m_i \left\| \widehat{e_i^{\mathcal{M}}} - e_i \right\|_2^2$$

where $\widehat{e_i^{\mathcal{M}}} = \text{Linear}(f(e^{\mathcal{M}}))$ represents the predicted embedding of a token e_i with $m_i = 1$, obtained by applying a linear transformation to the output hidden vector $f(e^{\mathcal{M}})$ from EEG Transformer encoder. This objective encourages the model to recover meaningful EEG representations from partial observations. The pretraining process is optimized using AdamW optimizer²⁸ with a learning rate scheduler and weight decay.

2.4 Task Heads

Task heads include 7 event-level task heads and 17 EEG-level task heads. The number of heads corresponds to the task types listed in Table 1 and differs between the two levels.

Specifically, the 7 event-level heads correspond to the following 7 tasks: normal vs. abnormal classification, sleep staging, burst suppression detection, slowing classification, spikes detection, spike localization, and seizure / IIIC classification. In each event-level task, a 10-second segment is assigned to only one class. These tasks include multiple classification objectives with varying output dimensions: normal vs. abnormal classification, burst suppression detection, and spikes detection are binary classification tasks, where the output is the probability after applying a sigmoid activation function; slowing classification and spike localization are both three-class classification tasks, where the outputs are three-dimensional vectors obtained after applying the softmax function. These vectors represent the probabilities for “others”, “focal slowing”, and “generalized slowing” in slowing classification, and for “others”, “focal spike”, and “generalized spike” in spike localization, respectively; seizure and IIIC classification are six-class classification tasks, outputting six-dimensional vectors of probabilities for the classes: “others”, “seizure”, “LPD”, “GPD”, “LRDA”, and “GRDA”; sleep staging is a five-class classification task, with the output being a five-dimensional vector representing the probabilities of “awake”, “N1”, “N2”, “N3”, and “REM”.

Unlike event-level tasks, all EEG-level tasks are binary classification tasks, as an entire EEG can simultaneously contain multiple labels from an event-level task. Therefore, all findings involved in event-level tasks are converted into separate binary classification tasks at the EEG level. Here, we implement 17 EEG-level tasks. The input to the EEG-level heads consists of the continuous outputs from the event-level heads, with the input dimensions determined by the number of classes in the corresponding event-level tasks. For example, as shown in Figure 3, the input to the spikes EEG-level head is a one-dimensional output from the spike detection event-level head, while the inputs to the EEG-level heads for seizure, LPD, GPD, LRDA, and GRDA all come from the six-dimensional output of the seizure and IIIC event-level head.

The event-level task-specific heads are directly built on top of the foundation model. We initialize the model with pretrained weights from the tokenizer and EEG Transformer encoder modules. During fine-tuning, only the last three Transformer encoder blocks and the linear classification head are updated, while the earlier layers remain frozen. No masking is applied during this stage, ensuring the model fully leverages the learned representations for downstream tasks. Among these tasks, since spikes are typically transient waveforms occurring within a 1-second window, shortening the segment length improves localization in continuous detection. Therefore, we perform spike detection on 1-second segments, i.e., ($w = 1s \times 200Hz$), whereas all other tasks are based on 10-second segments, as described previously, i.e., ($w = 10s \times 200Hz$).

EEG-level tasks require scanning the entire recording, which is impractical for Transformers due to their quadratic computational complexity. For instance, processing a 1-hour EEG would generate $(19 \times 60 \times 60 = 68,400)$ tokens, whereas models like GPT-4 typically support an input length of only 8,192 tokens (with some versions supporting up to 32K²⁹). This challenge becomes more pronounced when dealing with 24-hour EEGs. Therefore, our EEG-level task-specific heads are CNN-Transformer-based models independent of the foundation model. Their input consists of the continuous event-level outputs for the entire EEG, processed with a sliding step size of 1 second. Specifically, for an EEG recording of length L , the resulting sequence length is given by $\left\lceil \frac{\frac{L}{200\text{Hz}} - 10\text{s}}{1\text{s}} \right\rceil + 1$. The number of channels corresponds to the number of event-level classification tasks. The continuous event-level outputs pass through two convolutional layers and pooling, reducing the sequence length by 30, with each patch representing a 30-second EEG segment. A 2-layer Transformer block further processes the segments, yielding a binary classification result for each class. We set the absolute positional encoding with a maximum length of 15,000, allowing the model to support EEGs up to $\frac{15,000 \times 30}{60 \times 60} = 125$ hours which can accommodate all EEGs across our datasets.

Due to class imbalance and the varying complexity of EEG findings, we designed a training strategy that combines a focal loss³⁰ and curriculum learning^{31,32} to optimize the model for specific tasks:

$$\mathcal{L}_{\mathcal{H}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C \alpha_c \beta_i (1 - p_{i,c})^{\gamma} y_{i,c} \log p_{i,c} + \lambda (\beta_i - 1)^2$$

where C is the number of classes, $p_{i,c}$ is the softmax probability of class c for sample i , and $y_{i,c}$ is the one-hot encoded ground truth. The weighting factor α_c addresses class imbalance with $\sum_c \alpha_c = 1$, while the focusing parameter γ down-weights well-classified examples, emphasizing harder samples. β_i is a coefficient of the loss for each sample i , which is learned via gradient descent $\beta_i \leftarrow \eta \frac{\partial \mathcal{L}_{\mathcal{H}}}{\partial \beta_i} \cdot \frac{\partial \mathcal{L}}{\partial \beta_i} = \mathcal{L}_i > 0$; this implies that hard-to-classify instances with larger $\mathcal{L}_i$ result in a decrease in β_i . We initialize $\beta_i = 1$ and incorporate a regularization term to prevent rapid decay to zero.

3. Baselines

We evaluated MORGOTH against state-of-the-art (SOTA) models for different tasks. At the event-level, for spike detection, we compare MORGOTH with SpikeNet⁵ on SN2 and HEP datasets; for seizure and IIC classification, we compare MORGOTH with SPaRCNet⁶ and the winning entry from the Kaggle ‘‘Harmful Brain Activity Classification’’ competition³³ on IIC

dataset; For burst suppression detection, we compare MORGOTH with a published burst suppression detector³⁴ on MoE dataset. For sleep staging, we compare MORGOTH with U-Sleep model³⁵. In addition, U-Sleep requires a single EEG and a single EOG channel with arbitrary standard electrode placement, while UPenn PSG dataset has two EEG channels (C3, C4) and one EOG channel, so we compare MORGOTH with U-Sleep’s two results, referred to as U-Sleep1 and U-Sleep2. We also compare MORGOTH with other EEG foundation models LaBraM¹⁸, EEGFormer³⁶, BrainBERT³⁷ and EEG analysis baselines^{38,39} on TUH dataset with four predefined tasks for this dataset. At EEG level, we compared MORGOTH with SCORE-AI¹¹ on normal vs. abnormal, focal spike, generalized spike, focal slowing and generalized slowing tasks using SAI dataset. To assess whether the model matches or exceeds human-expert performance, we also compared MORGOTH’s performance against multiple experts.

4. Statistical Analyses

For evaluating the model accuracy, we report both Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves. When comparing performance with a given expert, we compute ROC and PR curves using the majority voting (excluding that expert’s votes, i.e., leave-one-expert-out ground truth) as the ground truth.

ROC curves were generated by varying the decision threshold across the full range of predicted probabilities (0 to 1). At each threshold, the true positive rate $TPR = \frac{TP}{TP + FN}$ and false positive rate $FPR = \frac{FP}{FP + TN}$ were calculated to capture the trade-off between sensitivity and the rate of false alarms. These points were then plotted to produce the ROC curve, providing a comprehensive view of the model’s discrimination performance across all possible operating points. For multi-class ROC analysis, we employed a one-vs-rest (OvR) strategy, treating each class as the positive class against all others combined as the negative class. ROC curves and AUCs were computed for each class individually and then averaged to provide an overall assessment of model performance across all classes.

PR curves were generated by varying the decision threshold across predicted probabilities. At each threshold, precision = $\frac{TP}{TP + FP}$ and recall = $\frac{TP}{TP + FN}$ were calculated to characterize the trade-off between positive predictive value and sensitivity. For multi-class problems, a one-vs-rest approach was used, treating each class as positive against all others, with macro-averaged AUC-PR reported to summarize overall performance. For multi-class PR analysis, we employed a one-vs-rest strategy, treating each class as the positive class against all others combined as the negative class. Precision and recall were computed across a range of decision thresholds to characterize the trade-off between positive predictive value and sensitivity. PR curves were generated for each class, and

the area under the precision-recall curve (AUC-PR) was macro-averaged to provide an overall assessment of model performance across all classes.

Confusion Matrix for Binary Classification is:

	Predicted Positive	Predicted Negative
Actual Positive	TP (True Positive)	FN (False Negative)
Actual Negative	FP (False Positive)	TN (True Negative)

For the classification evaluation methods used in each task, see Table S1. We employed binary classification evaluation for the binary tasks (17 EEG-level tasks; event-level abnormal detection, burst suppression detection, and spike detection). For multi-class tasks (event-level sleep staging, slowing detection, spike localization, and IIC classification), we used a one-vs-rest approach, evaluating each feature or class separately. Overall performance for these multi-class tasks was then summarized by averaging the results across all classes.

We also calculate the Experts Under the Curve (EUC) metric, the percentage of experts' ROC or PR operating points that fall below the model's corresponding curve. EUC measures the degree to which a model matches or exceeds expert performance. For all performance metrics, we calculated 95% confidence intervals (CIs) and p-values through 10,000 rounds of bootstrapping. Specifically, we computed the Experts Under the Curve (EUC) metric as follows: for each classification task, we first generated the model's ROC or PR curve by varying the decision threshold. Each expert's performance on the same task was summarized as a single operating point (i.e., their sensitivity and false positive rate (1-specificity) on the ROC space, or precision and recall on the PR space). We then determined whether each expert's operating point lay below the model's curve—meaning the model achieved equal or higher sensitivity (or precision) at the same or lower false positive rate (or recall). The EUC was finally calculated as the percentage of experts whose operating points fell below the model's curve. This provides an intuitive summary of how frequently the model outperformed individual experts across the spectrum of decision thresholds.

To assess the reliability of the results, we report Inter-Rater Reliability (IRR) based on Cohen's κ . We compute pairwise IRR among all pairs of experts (expert-expert κ), expert-Morgoth and expert-SOTA pairs (expert-algorithm κ), and majority IRR between each of these (expert or algorithm) and the majority consensus of human experts. Pairwise IRR analysis, measured using Cohen's κ , is limited to experts who labeled at least 1,000 samples in common per class for event-level tasks and at least 30 patients in common per class for EEG-level tasks. We adopt the convention of describing ranges of by κ values as indicating "slight" (0% – 20%), "fair" (21% – 40%), "moderate" (41% – 60%), "substantial" (61% – 80%), and "almost perfect" (81%-100%) reliability⁴⁰. We compared experts' and algorithms' statistical calibration for each class using soft scores derived from multi-experts. For each expert or algorithm, we estimated their probability of classifying segments within each of these probability bins as that by fitting a parametric model⁸.

We further define a calibration index ranging from -1 (maximal under-calling) to 1 (maximal over-calling), with 0% representing perfect calibration. In addition, we define significant over-calling and under-calling as having a calibration index above 20% or below -20%.

After evaluating the calibration, we used Platt scaling and isotonic regression to calibrate model outputs to better reflect true probabilities.

Platt scaling is a parametric probability calibration technique that fits a sigmoid function to map the raw output scores of a classifier (such as the decision values from an SVM or the logits from a neural network) into calibrated probabilities. The key assumption is that the relationship between the uncalibrated scores and the true probabilities can be well approximated by a logistic (sigmoid) curve. Formally, given the original model scores $f(x)$, Platt scaling fits a sigmoid of the form:

$$P(y = 1 | x) = \frac{1}{1 + \exp(A f(x) + B)}$$

where the parameters A and B are learned by minimizing the negative log-likelihood (cross-entropy loss) on a calibration set. Specifically, if we have calibration samples (f_i, y_i) where $y_i \in \{0,1\}$, we solve:

$$\min_{A,B} - \sum_i (y_i \log p_i + (1 - y_i) \log (1 - p_i))$$

$$p_i = \frac{1}{1 + \exp(A f_i + B)}$$

This optimization is typically performed using a standard numerical solver such as Newton's method.

Isotonic regression is a non-parametric calibration technique that maps classifier scores to calibrated probabilities under the constraint that the mapping is monotonically increasing. Given a calibration set $\{(f_i, y_i)\}_{i=1}^N$, where f_i are the model output scores and $y_i \in \{0,1\}$ are the corresponding binary labels, isotonic regression seeks to find a function $m(\cdot)$ that solves:

$$\min_{\{m\}} \sum_{i=1}^N (y_i - m(f_i))^2$$

subject to the monotonicity constraint:

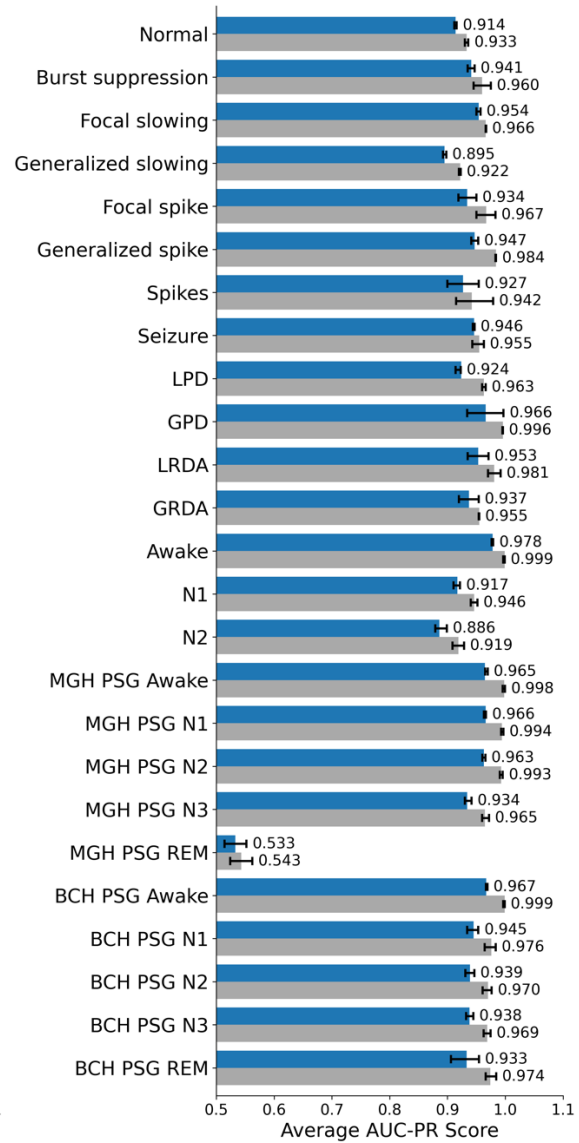
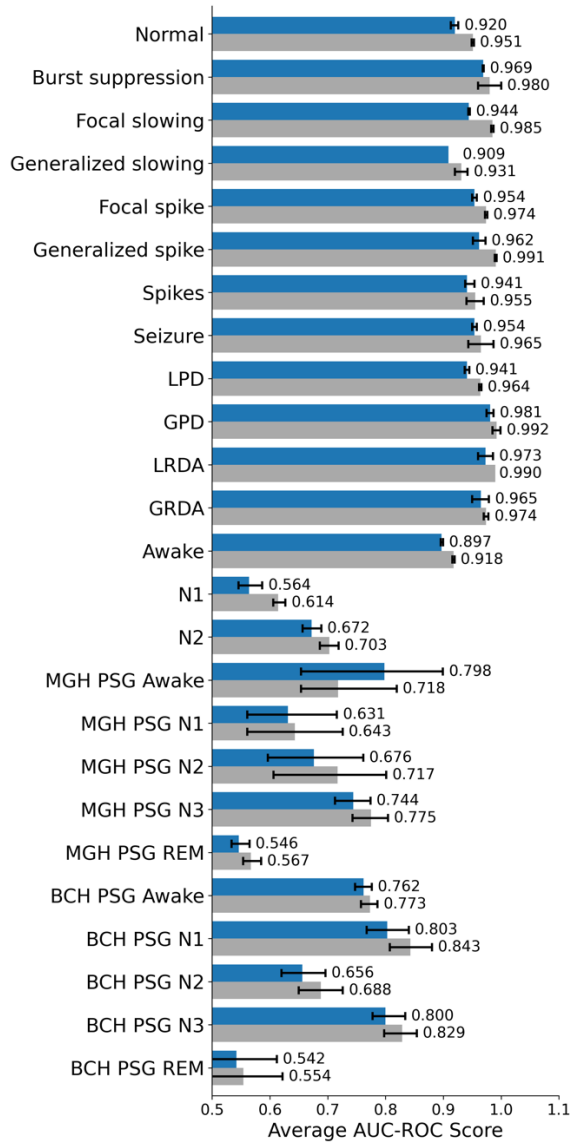
$$m(f_i) \leq m(f_j) \text{ whenever } f_i < f_j.$$

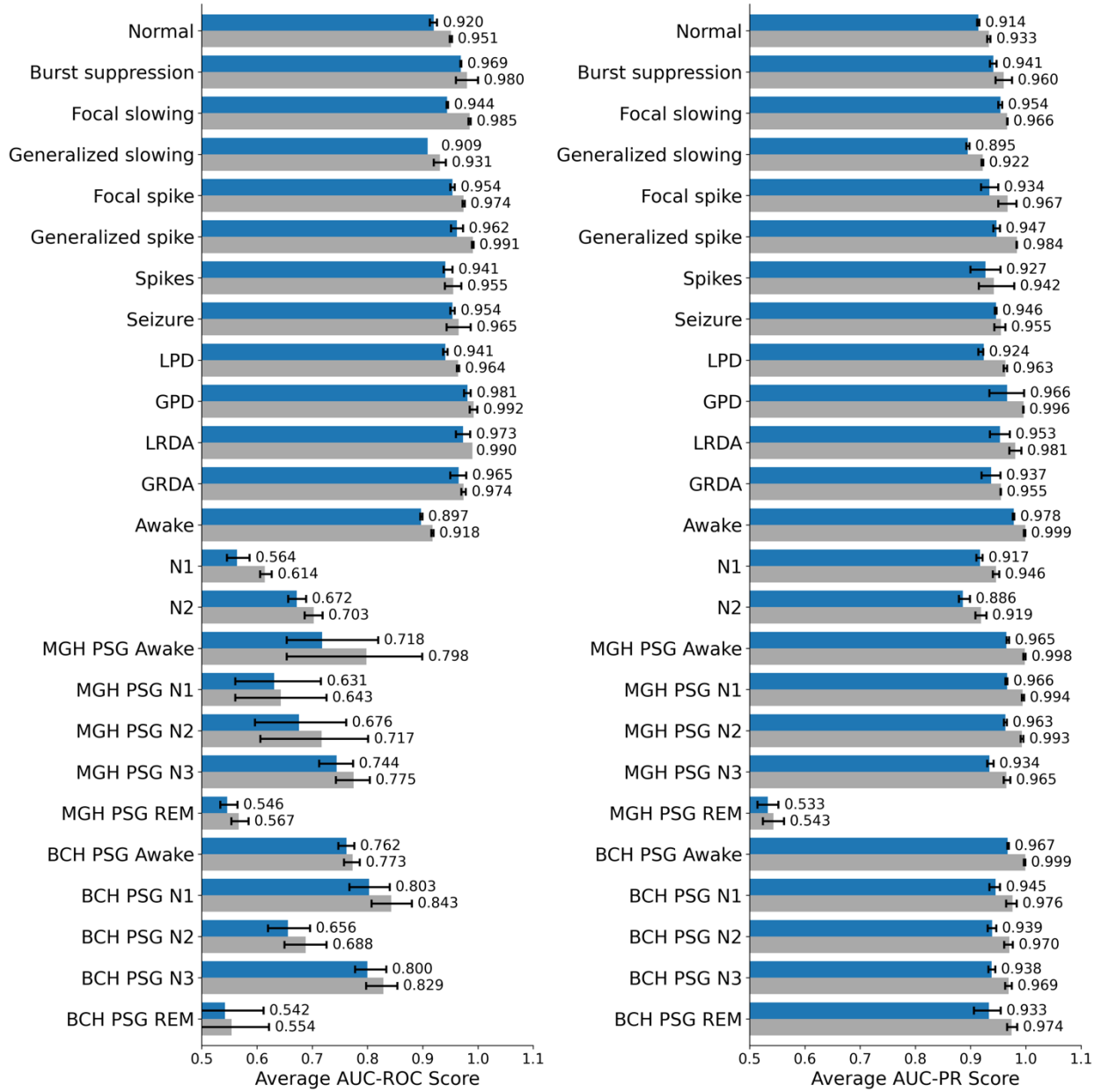
The solution is typically a piecewise constant, non-decreasing function (i.e., a step function) that best fits the empirical data while respecting the monotonic relationship.

5. Results for Single-scored Validation Datasets

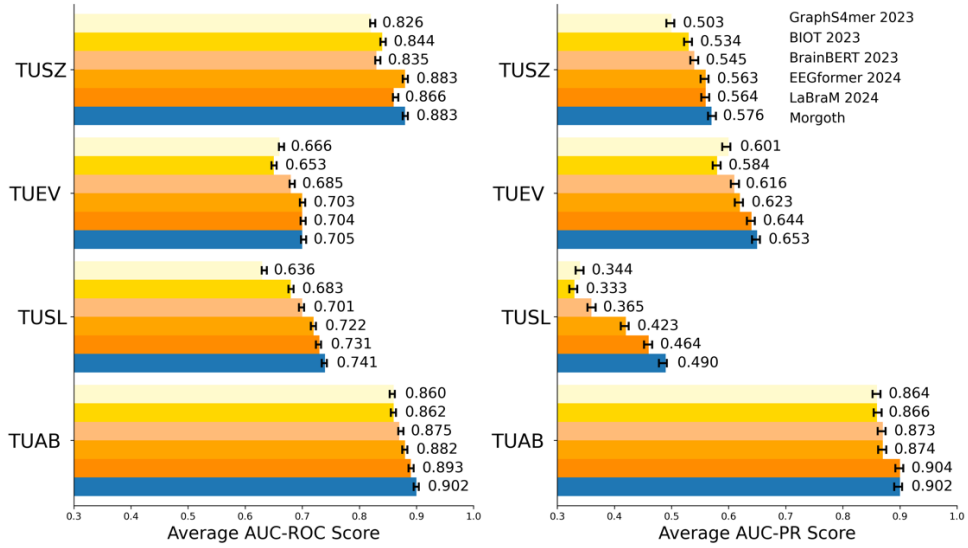
On internal test datasets with labels from a single expert (Figure S4A), MORGOTH demonstrated high accuracy, achieving an AUC-ROC of 0.98 and 0.92 for distinguishing normal vs. abnormal EEG patterns at both the event and EEG levels, respectively; 0.98 and 0.94 for focal slowing; 0.98 and 0.90 for generalized slowing; 0.88 and 0.97 for burst suppression; 0.97 and 0.89 for spikes; 0.90 and 0.95 for focal spikes; 0.93 and 0.96 for generalized spikes; 0.86 and 0.95 for seizure; 0.90 and 0.94 for LPD; 0.91 and 0.98 for GPD; 0.91 and 0.97 for LRDA; 0.90 and 0.96 for GRDA. For sleep staging, AUC-ROC values were 0.92, 0.72, 0.95, 0.89, and 0.87 across the five sleep stages (awake, N1, N2, N3, REM), comparable with reported accuracies of experts judged against scores of other experts⁴¹. On internal HEEDB database, the difference between the training and test accuracy averaged 2.75%, with the test accuracy not falling more than 5% below the training accuracy, indicating that model overfitting was minimal.

Performance on external datasets scored by a single expert is shown in Figure S4B-C. On TUH dataset benchmark, MORGOTH achieves equivalent or superior performance values across all four tasks (Figure S4B). On MASS sleep staging task, MORGOTH outperforms U-Sleep (Figure S4C).

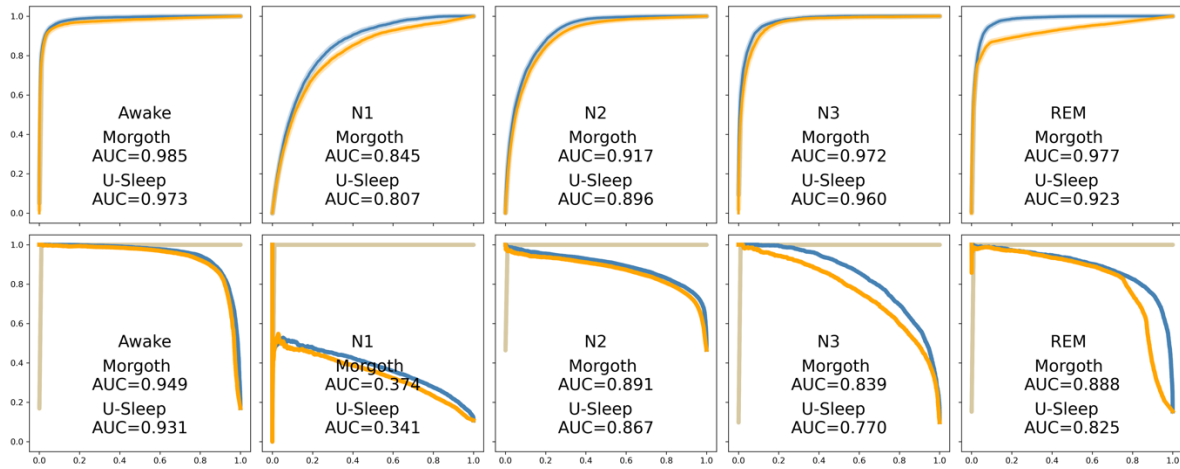




(A) Internal Dataset: Morgoth's average AUC-ROC scores (left) and AUC-PR scores (right) with 95% CIs for event-level tasks (first) and EEG-level tasks (second) on internal HEEDB-Train, MGH-PSG-Train sets (gray) and HEEDB-Test, MGH-PSG-Test, and BCH-PSG-Test sets (blue). The PSG data from BCH was not used in training, so there is no gray bar.



(B) External datasets: Average AUC-ROC scores (left) and AUC-PR scores (right) with 95% CIs of Morgoth (blue) and baselines (yellows) on external TUH dataset, which has tasks TUAB (normal/abnormal), TUSL (seizure/slowing/background), TUSZ (seizure/not seizure), and TUEV (spike and slow wave/generalized periodic epileptiform discharge/periodic lateralized epileptiform discharge/eye movement/artifact/background).



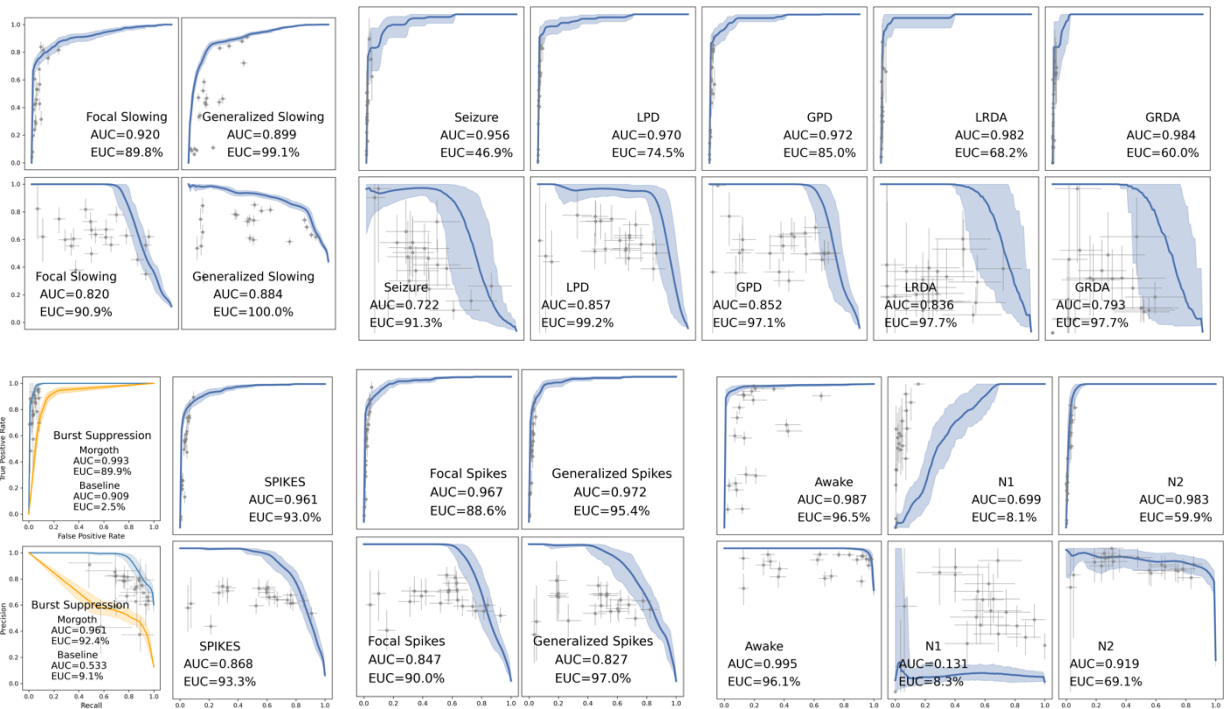
(C) External datasets: Morgoth's ROC curves (top row) and PR curves (bottom row) on external MASS dataset with 95% CIs.

Figure S4. Performance Results on Single-scored Datasets.

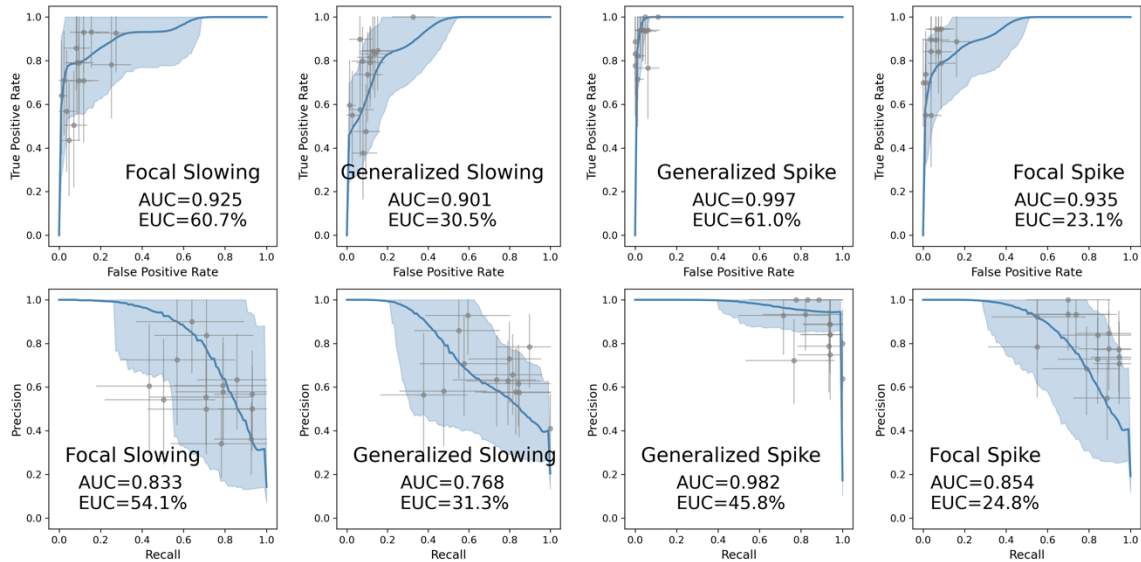
6. Results for Multi-expert Validation Datasets

Figure S5A presents MORGOTH's performance across 14 EEG findings on the internal MoE test dataset, MORGOTH achieved an average AUC-ROC of 0.944 with a 95% CI [0.936, 0.952], and an average AUC-PR of 0.800 [0.770, 0.826] across 14 categories. Comparing ROC and PR curves with the operating points of 21 experts, MORGOTH outperformed an average of 90% of experts (19 out of 21, range 75.2% to 100%) across 13 of the 14 tasks, except for sleep stage N1.

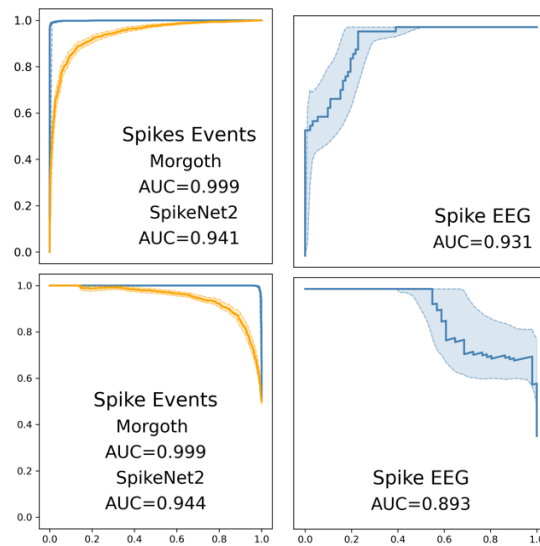
Figure S5B shows 4 EEG-level tasks results on external ON dataset. In comparison with the results of 15 experts, MORGOTH achieves an average EUC-ROC of 51.5% [46.18%, 56.18%] and an average EUC-PR of 39.3% [25.8%, 57.5%], with an AUC-ROC of 0.939 [0.936, 0.943] and an AUC-PR of 0.857 [0.848, 0.870].



(A) Internal MoE dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue), burst suppression baseline (yellow), with 95% CIs, and expert points (gray).



(B) External ON dataset: Average ROC curves (odd rows) and PR curves (even rows) of Morgoth (blue) with 95% CIs, and expert points (gray)



(C) External HEP dataset: Average event-level (left) and EEG-level (right) spike detection ROC curves (odd rows) and PR curves (even rows) of Morgoth (blue) and SpikeNet 2 (yellow), with 95% CIs. Since the experts assigned were not the same across all EEGs, it was not possible to plot individual expert points or calculate EUC performance.

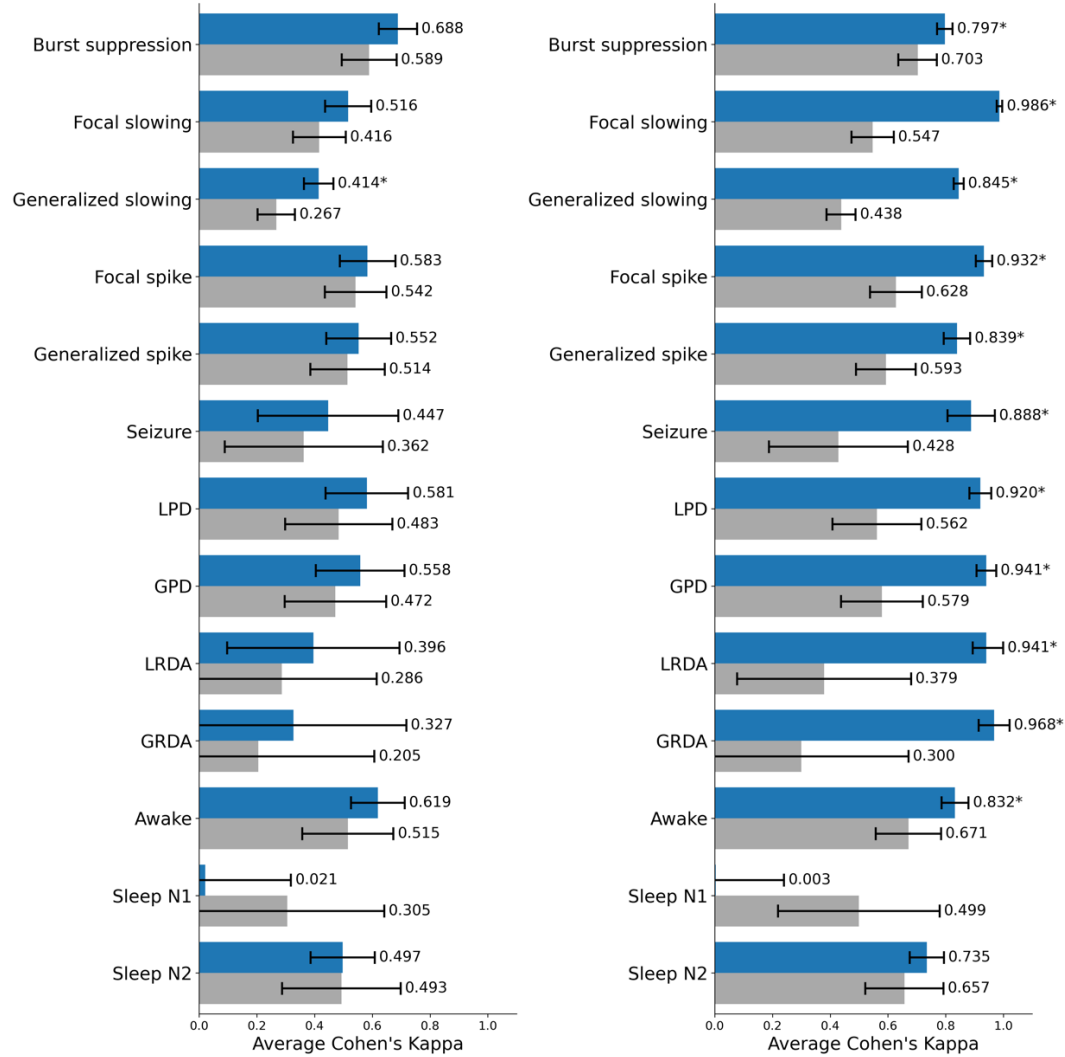
Figure S5. Additional Performance Results on Multi-expert Scored Datasets.

7. Inter-rater Reliability

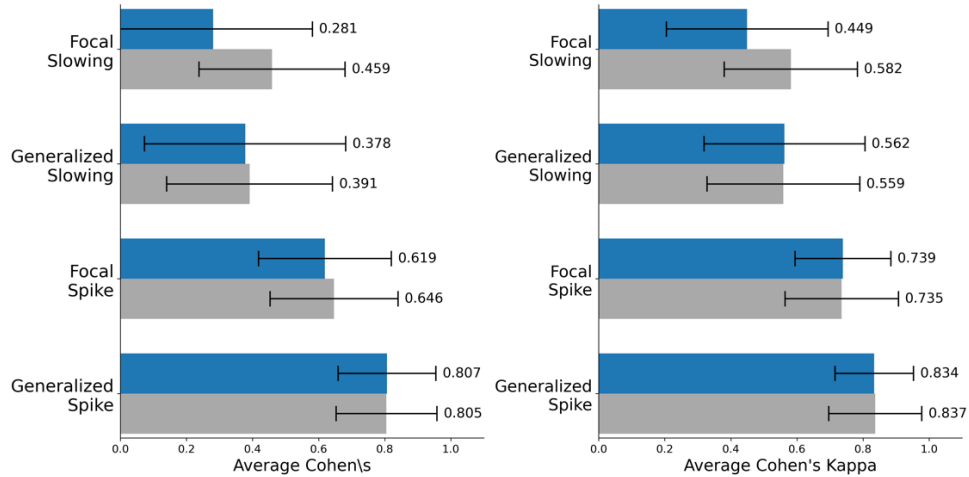
To validate both the reliability of the model and the consistency of expert annotations, we evaluated the agreement of each expert and the model with individual experts as well as the agreement with consensus across all multi-expert annotated datasets. Figures 4 and S6 provide a comprehensive analysis, presenting the average κ values along with 95% CI bars. Pairwise IRR analysis is limited to experts who labeled at least 1,000 samples in common per class for event-level tasks, at least 30 patients in common per class for EEG level classes. We obtained pairwise IRRs among 21 experts with 394 expert pairs in MoE dataset, 30 experts with 772 expert pairs in IIIC-Test dataset, 8 experts with 56 expert pairs in SN2-Test dataset, 14 experts with 164 expert pairs in SAI dataset, 14 experts with 182 expert pairs in ON dataset, and 6 experts with 30 expert pairs in UPenn dataset.

In MoE dataset, expert pairwise IRR is “fair” for generalized slowing, seizure, LRDA, GRDA, N1, “moderate” for burst suppression, focal slowing, focal spike, LPD, GPD, awake and N2. While MORGOTH achieved “substantial” for burst suppression and awake, “fair” for generalized slowing and seizure, surpassing expert performance in κ scores across all categories except for N1. Similarly, the majority IRR also reached at least a “fair” level across all categories, with overall scores higher than pairwise κ . MORGOTH consistently outperformed experts in this evaluation, achieving higher κ scores in all categories except for N1. In the IIIC-Test dataset, expert pairwise IRR was “fair” for seizure, LRDA, and GRDA and “moderate” for LPD and GPD, while expert majority IRR was “fair” for LRDA and “moderate” for the other four categories. MORGOTH achieved higher κ scores across both metrics, with a statistically significant improvement in LPD. In the SN2-Test dataset, expert pairwise IRR was “fair”, while majority IRR reached a “moderate” level. MORGOTH achieved “moderate” and near “perfect” agreement, with a statistically significant improvement over expert performance. In SAI dataset, expert pairwise IRR was “moderate” for focal slowing, generalized slowing, and focal spike, and “substantial” for normal and generalized spike. Majority IRR was “moderate” for generalized slowing and “substantial” for the other four categories. Although MORGOTH did not show a significant improvement in agreement, it consistently maintained the same levels of reliability as experts. In ON dataset, MORGOTH and experts exhibited similar levels of pairwise IRR and majority IRR. Specifically, pairwise IRR was “fair” for generalized slowing, “moderate” for focal slowing, “substantial” for focal spike, and almost “perfect” for generalized spike. Majority IRR was “moderate” for both focal and generalized slowing, “substantial” for focal spike, and almost “perfect” for generalized spike. In UPenn dataset, MORGOTH and experts exhibited similar levels of pairwise IRR and majority IRR. Specifically, pairwise IRR was “slight” for N1, “moderate” for N2 and N3, and “substantial” for awake and REM. Majority IRR was “fair” for N1, “moderate” for N3, “substantial” for awake and N2, and almost “perfect” for REM. N1 showed low expert agreement, indicating its difficulty in classification and the potential unreliability of using expert labels as ground truth for model evaluation.

These results confirm the reliability of expert labels and the stability of group consensus as a reference for evaluation. The majority inter-rater reliability (IRR), consistently surpassing pairwise κ , highlights its reproducibility. Meanwhile, MORGOTH’s strong agreement with both experts and consensus – often exceeding expert performance – demonstrates its robustness and generalization. This reinforces the effectiveness of multi-expert annotations for training and evaluation.



(A) MoE dataset



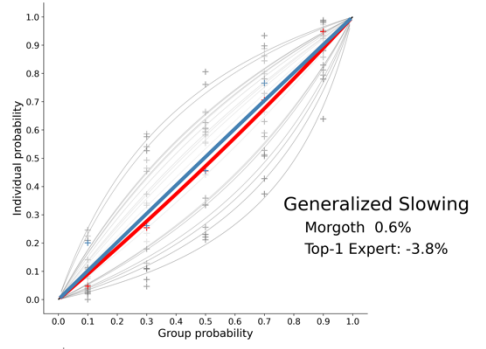
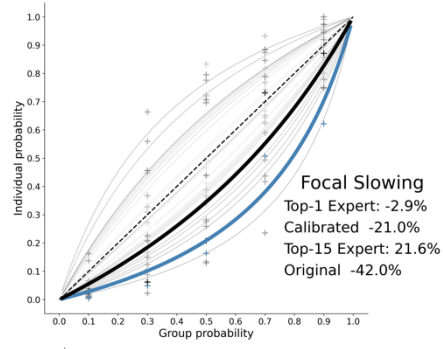
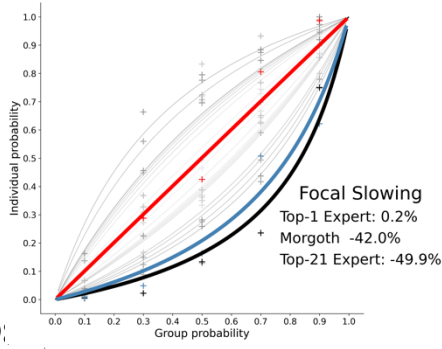
(B) External ON dataset

Figure S6. Additional Reliability Analyses. Inter-rater Agreements of Internal Dataset. Left: average agreement with 95% CIs across all pairwise experts (gray), and average agreement between Morgoth (blue) with each expert. Right: Average agreement with 95% CIs between Morgoth, baselines, or one expert with consensus from leaving one expert out.

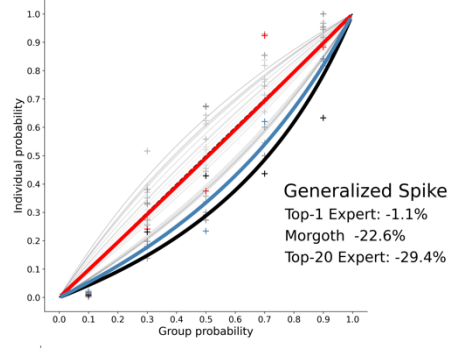
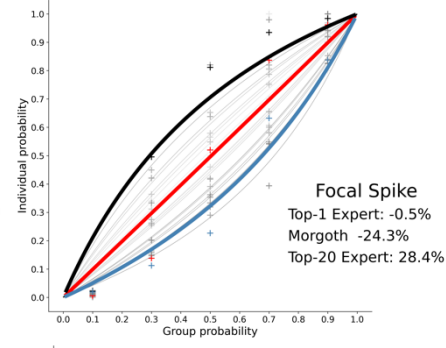
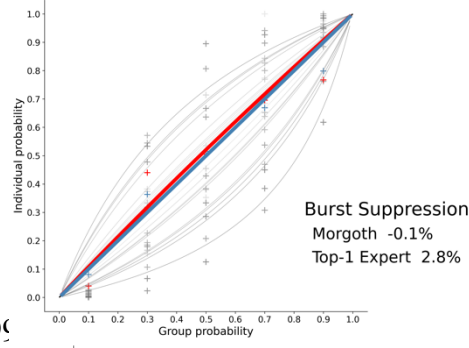
8. Calibration Analysis

Figure 4 and Figure S7 presents calibration curves for experts and MORGOTH across all categories in MoE, IIIC-Test, SN2-Test, SAI, ON, UPenn datasets. MORGOTH's calibration index falls within -20% to 20% for 10 out of 16 event level categories and 4 out of 5 EEG level categories, with absolute values ranging from 0.1% to 19.8, indicating near-perfect calibration, surpassing 40.0% to 96.7% of experts, and outperforming all three baseline models. Especially in MoE's generalized slowing, burst suppression, LPD, GRDA, and ON's focal slowing, it outperformed the best expert.

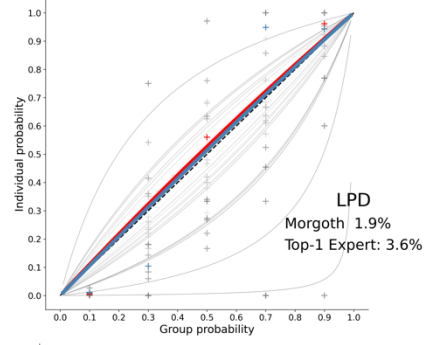
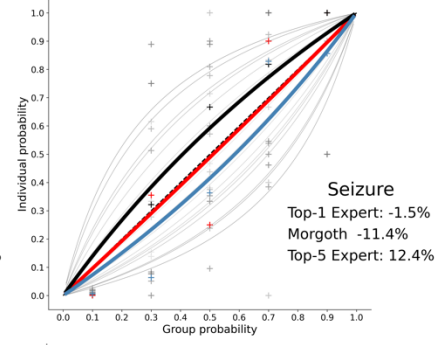
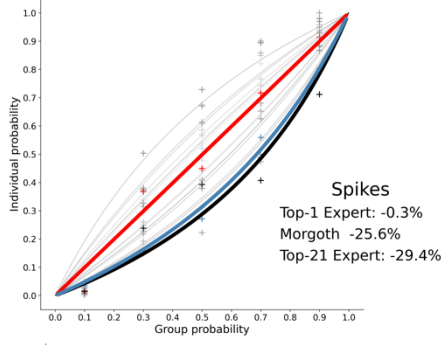
69



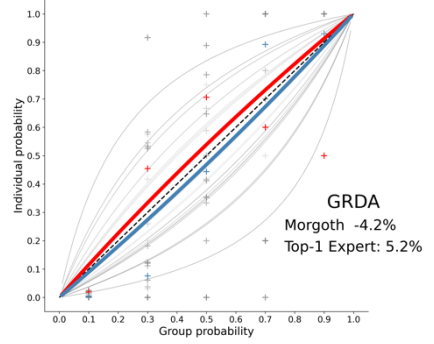
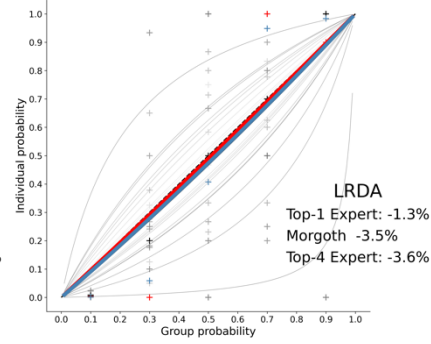
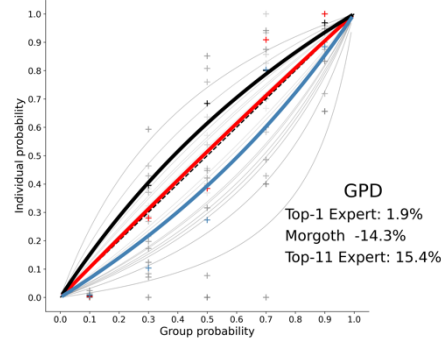
69

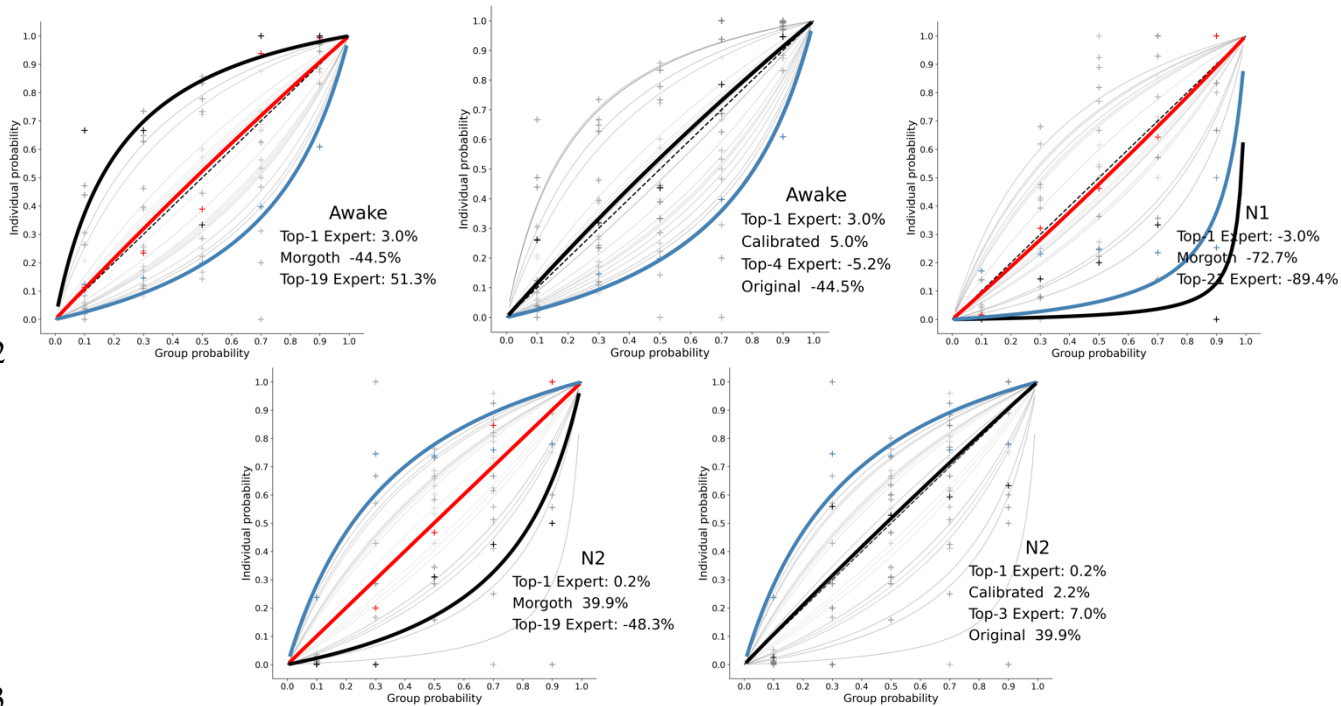


700

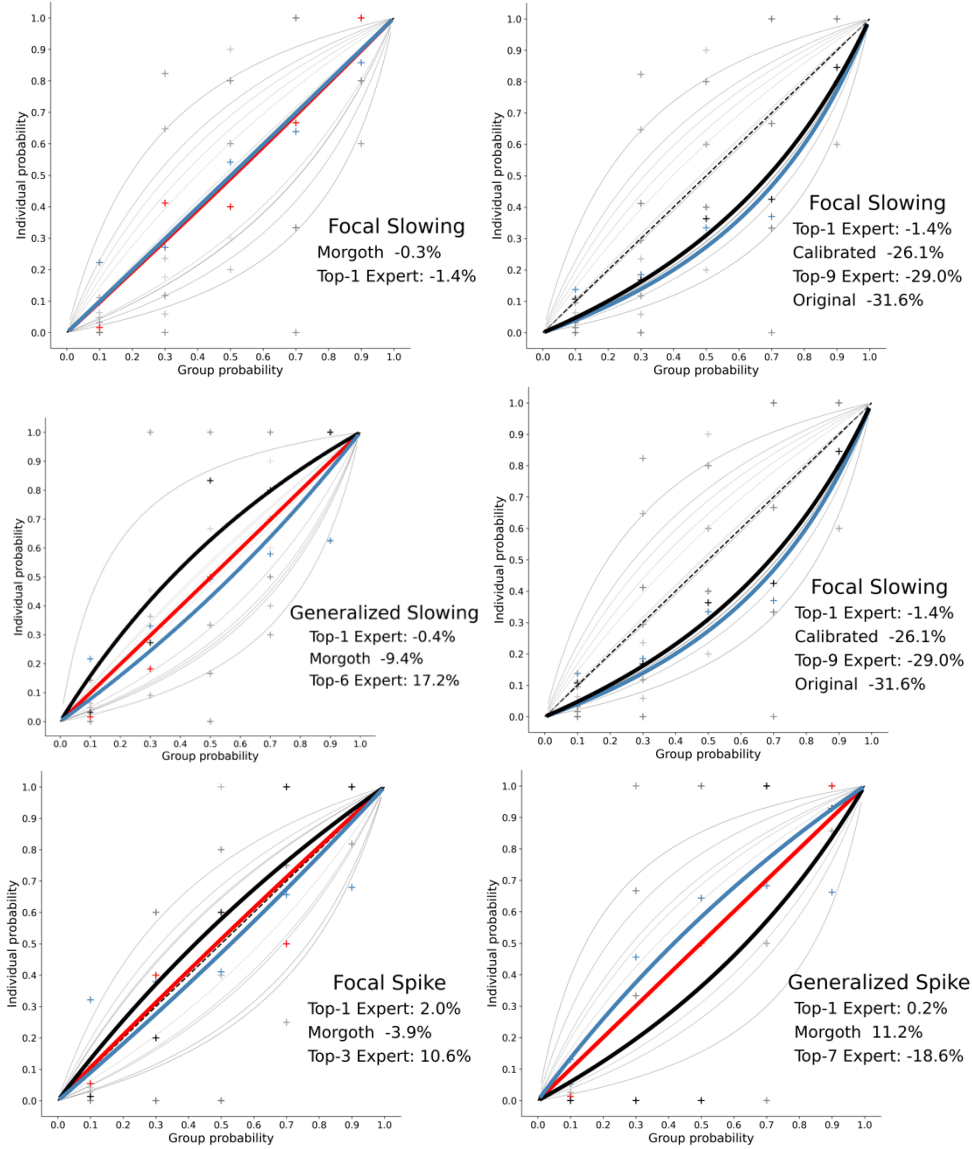


701





(A) MoE dataset



(B) ON dataset

Figure S7. Additional calibration curves for (A) MoE dataset and (B) ON dataset. In the plots, MORGOTH is shown in blue and experts in gray. The black line represents the best-performing expert, and the best expert who performs worse than the model.

9. Case Studies

Figures S8-S11 illustrate MORGOTH's continuous EEG interpretation for spike detection, including an example of 3Hz generalized spike and wave activity (Figure S8). Additionally, the figures demonstrate its capability in rejecting false positives and artifact ensemble spikes, such as vertex waves (Figure S9), sawtooth patterns (Figure S10), and ECG artifacts (Figure S11). Figures

[S12-S21](#) showcase MORGOTH's continuous EEG interpretation for seizure and IIC classification, including 2 seizures (Figures [S12](#) and [S13](#)), 2 LPD (Figures [S14](#) and [S15](#)), 2 GPD (Figures [S16](#) and [S17](#)), 2 LRDA (Figures [S18](#) and [S19](#)), and 2 GRDA (Figures [S20](#) and [S21](#)).

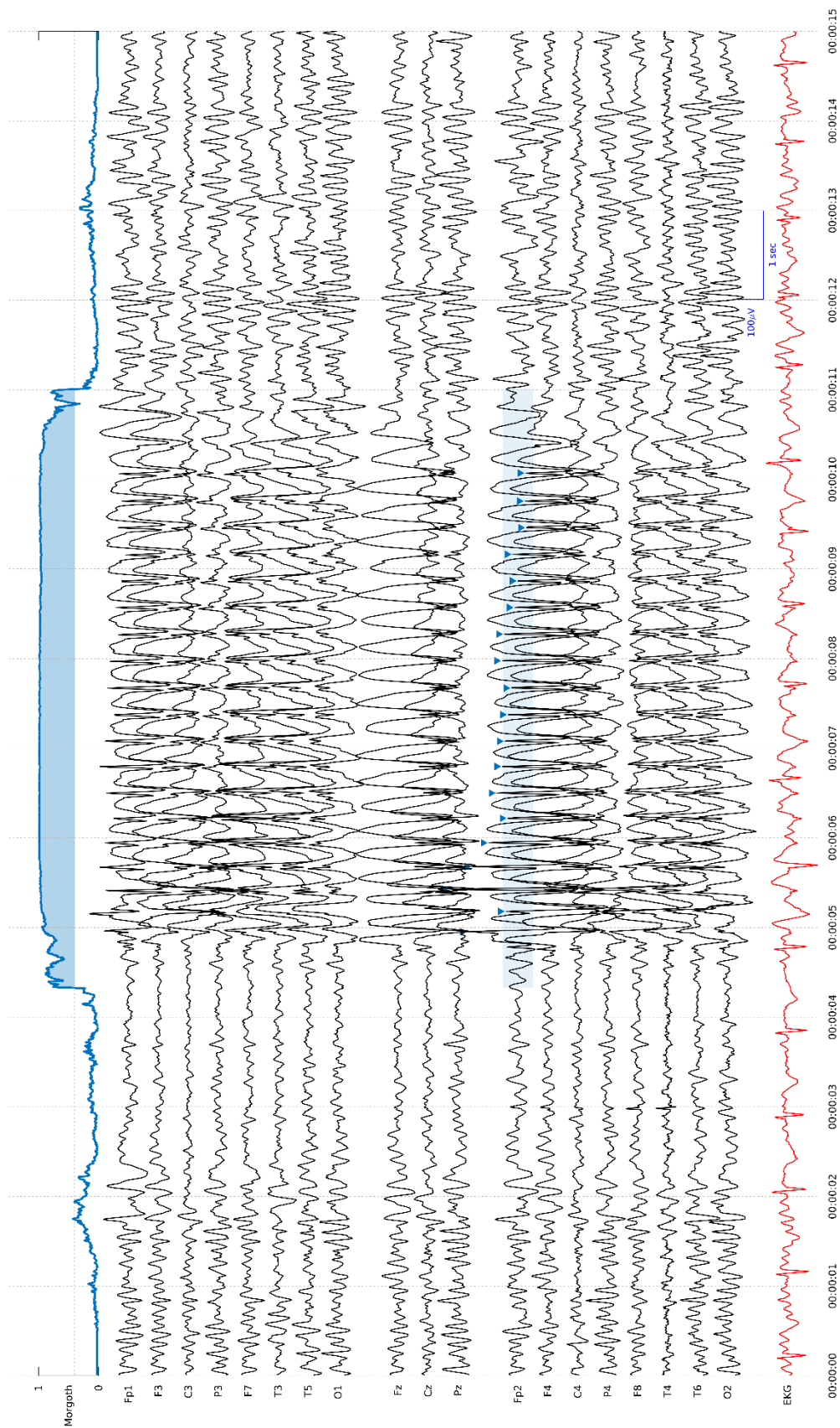


Figure S8. A 6-second run of 3 Hz generalized spike and wave activity detected by Morgoth.

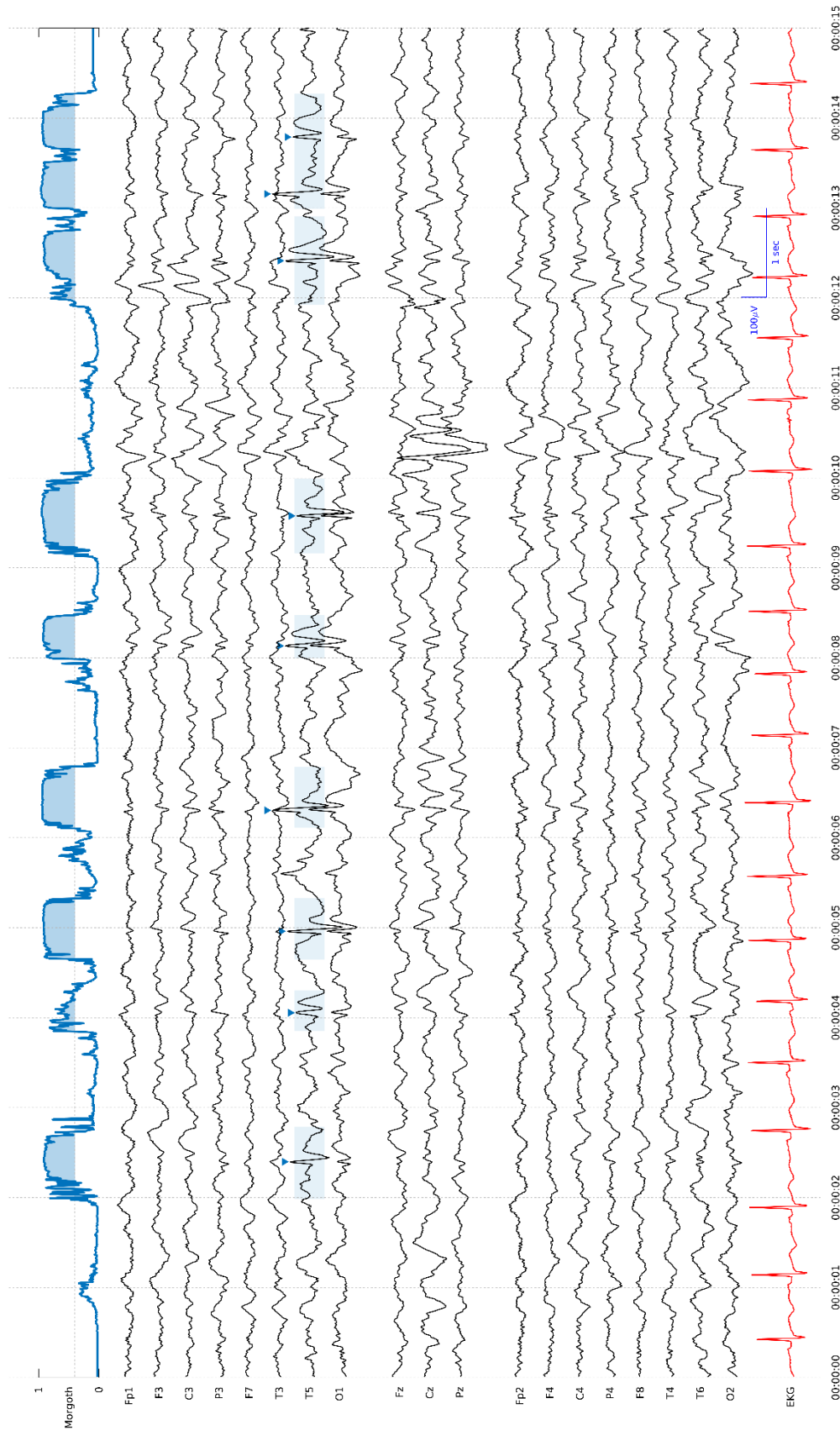


Figure S9. A series of left posterior temporal spike and wave discharges maximal at the T5 electrode detected by Mergoth. Approximately 10 seconds into this sample, there is a vertex wave maximal at CZ that resembles an epileptic discharge but is correctly rejected by Mergoth.

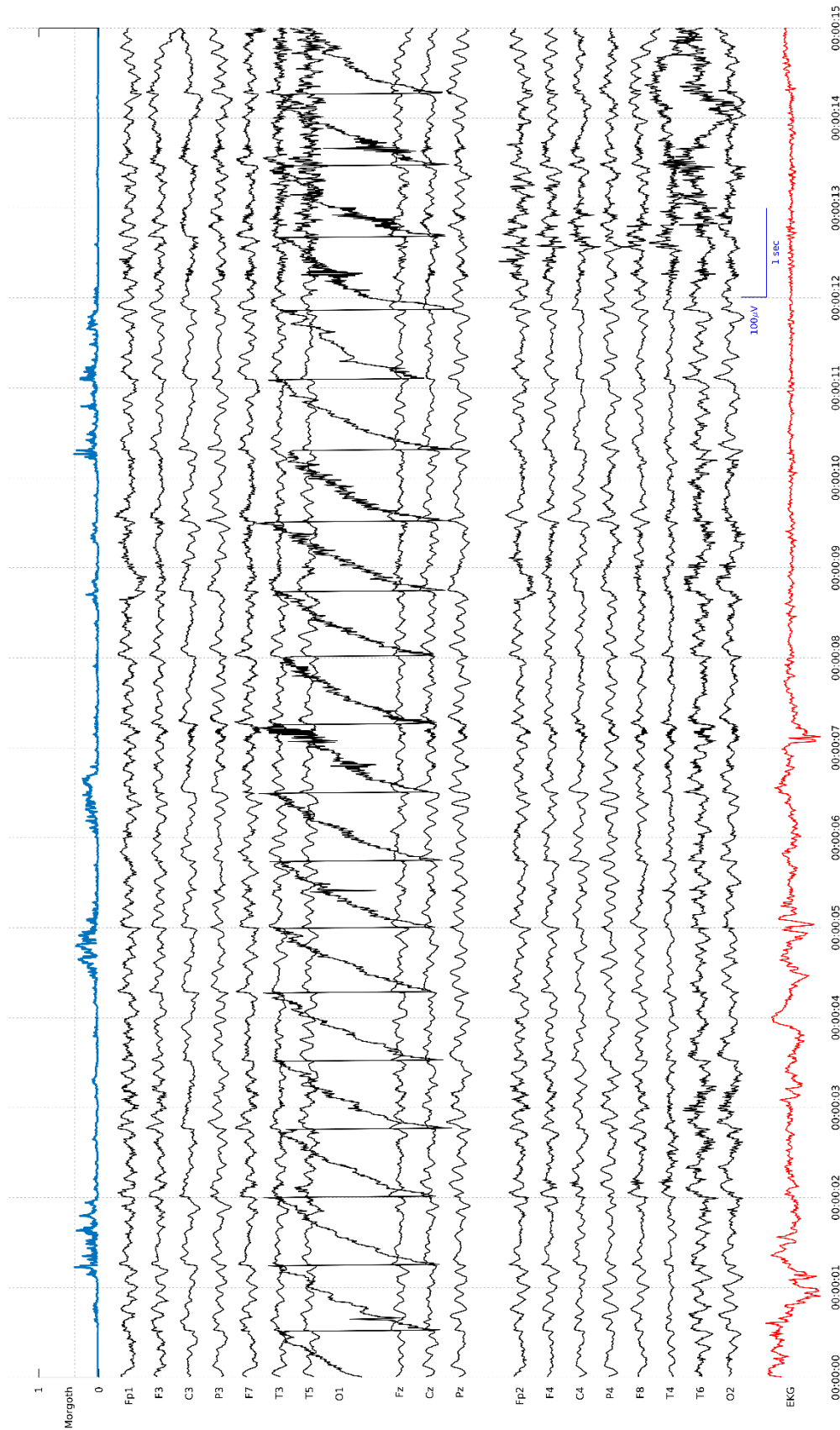


Figure S10. Sawtooth artifact produces high amplitude repetitive sharp triangular patterns. These are correctly rejected by Morghani as not being epileptiform discharges.

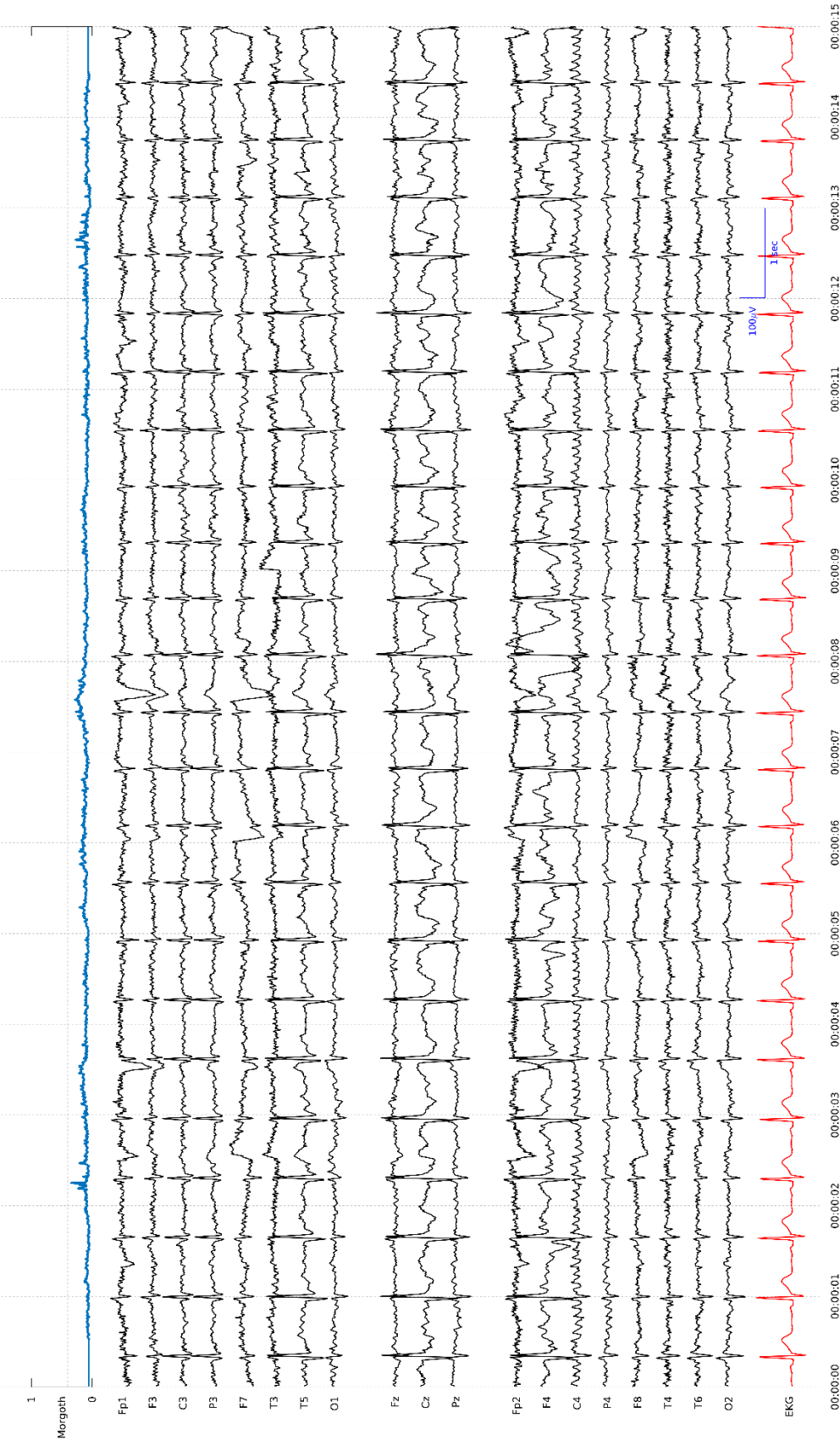


Figure S11. A series of epileptiform-appearing ECG artifacts that are correctly rejected by Morgoth as not being epileptiform discharges.

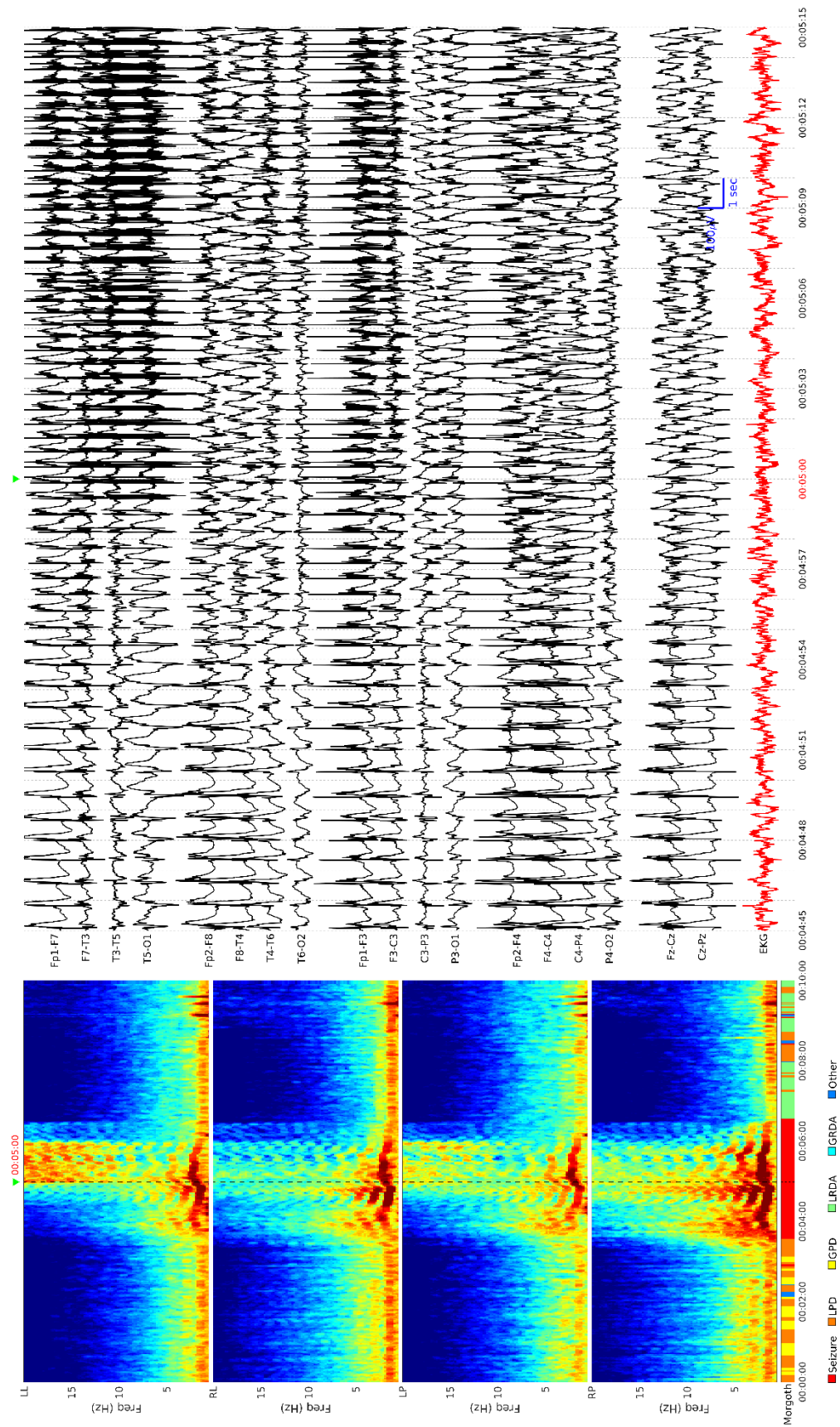


Figure S12. A generalized seizure detected by Morgoth, most prominent in the right parasagittal region, which in the spectrogram lasts approximately 2 minutes. The EEG sample corresponds to the green triangle and vertical dashed line, consisting of repetitive epileptic discharges with increasing frequency. Beneath the spectrogram, the color bar shows that Morgoth is detecting somewhat random-looking runs of other rhythmic and periodic IIC patterns, including LPDs (orange), GPDs (yellow), and LRDA (green).

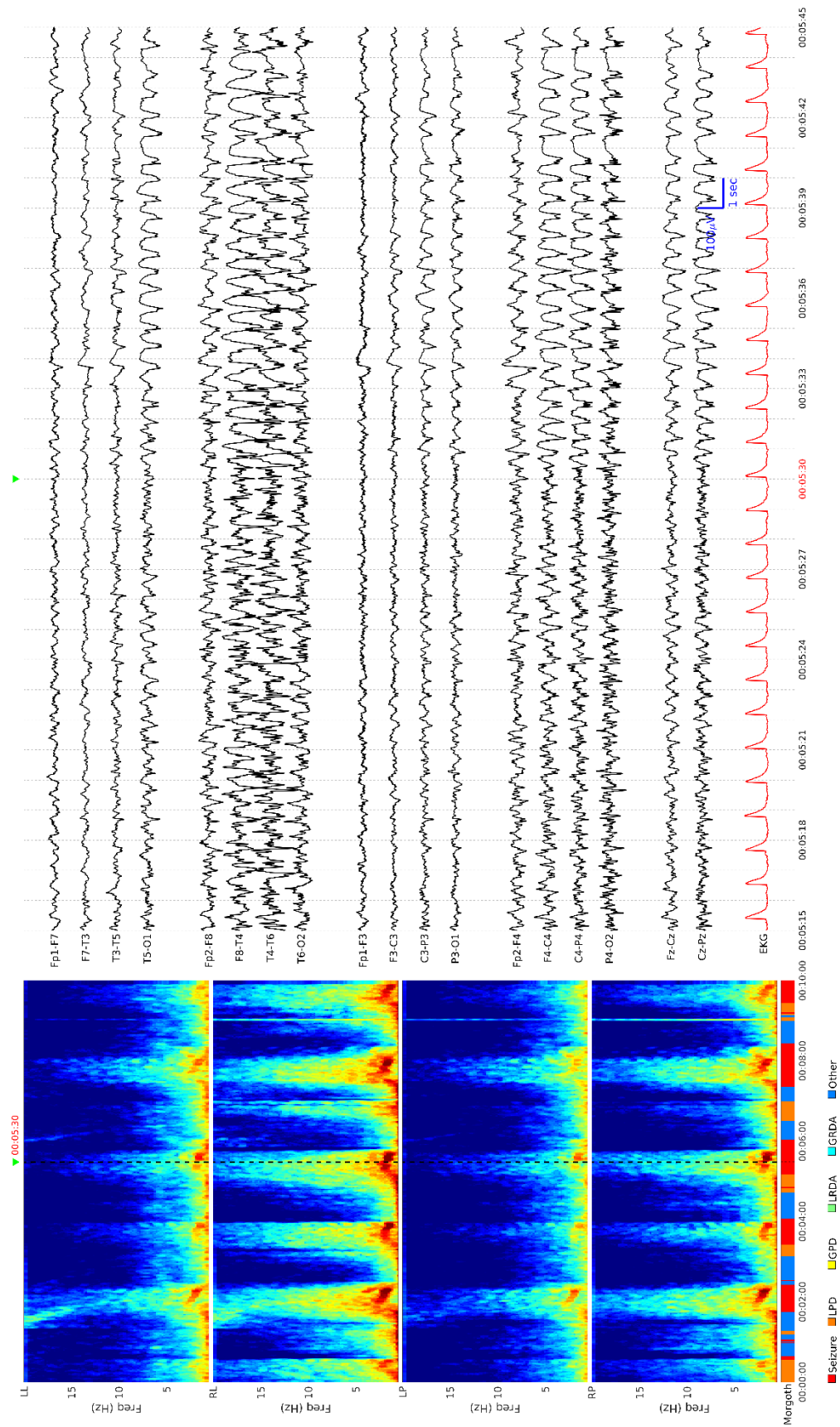


Figure S13. A series of seizures from the right hemisphere occurring at roughly 2-minute intervals, each lasting about 1 minute, detected by Morgoth. Beneath the spectrogram, Morgoth detects these seizures (red) alternating with LPDs (orange) and periods without IIIC patterns (blue). The EEG sample, corresponding to the green triangle and vertical dashed line, shows evolving rhythmic activity increasing in frequency.

Figure S14. Left posterior parietal-temporal Lateralized Periodic Discharges (LPDs) at approximately 0.5 Hertz correctly detected by Morgoth. The 10-minute spectrogram shows a left hemisphere broadband monotonous pattern consistent with the LPDs. The green triangular marker and vertical dashed line on the spectrogram indicate the location of the EEG sample on display. The orange bar along the bottom shows Morgoth's classifications of the pattern time-aligned with the spectrogram, indicating continuous LPDs.

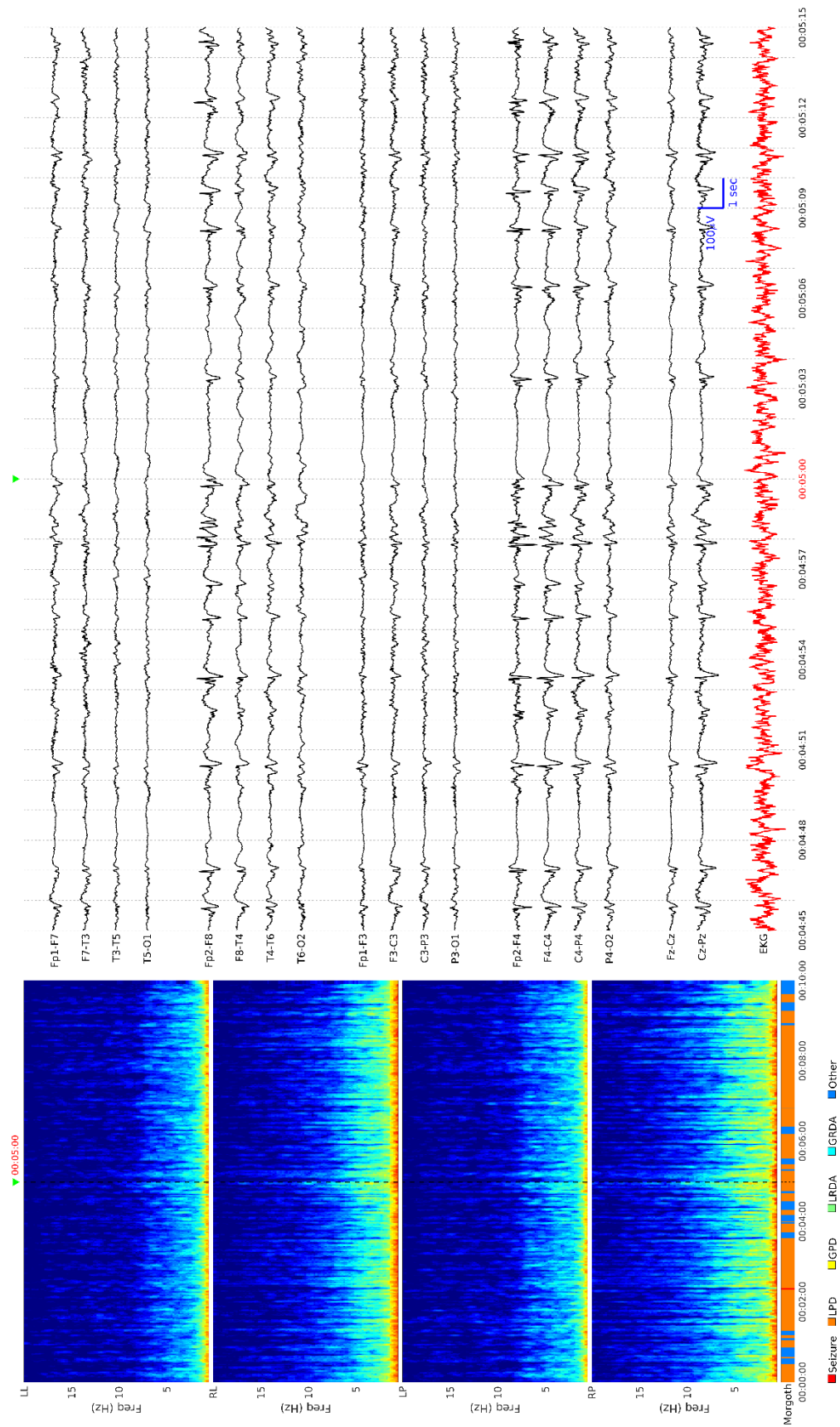


Figure S15. Right frontal-central predominant Lateralized Periodic Discharges (LPDs) at approximately 1Hz detected by Morgoth. The 10-minute spectrogram shows a left hemisphere broadband monotonous pattern consistent with LPDs. The green triangular marker and vertical dashed line show the location of the EEG sample. The orange bar along the bottom shows Morgoth's classifications time-aligned with the spectrogram, indicating LPDs (orange) which intermittently resolve (blue).

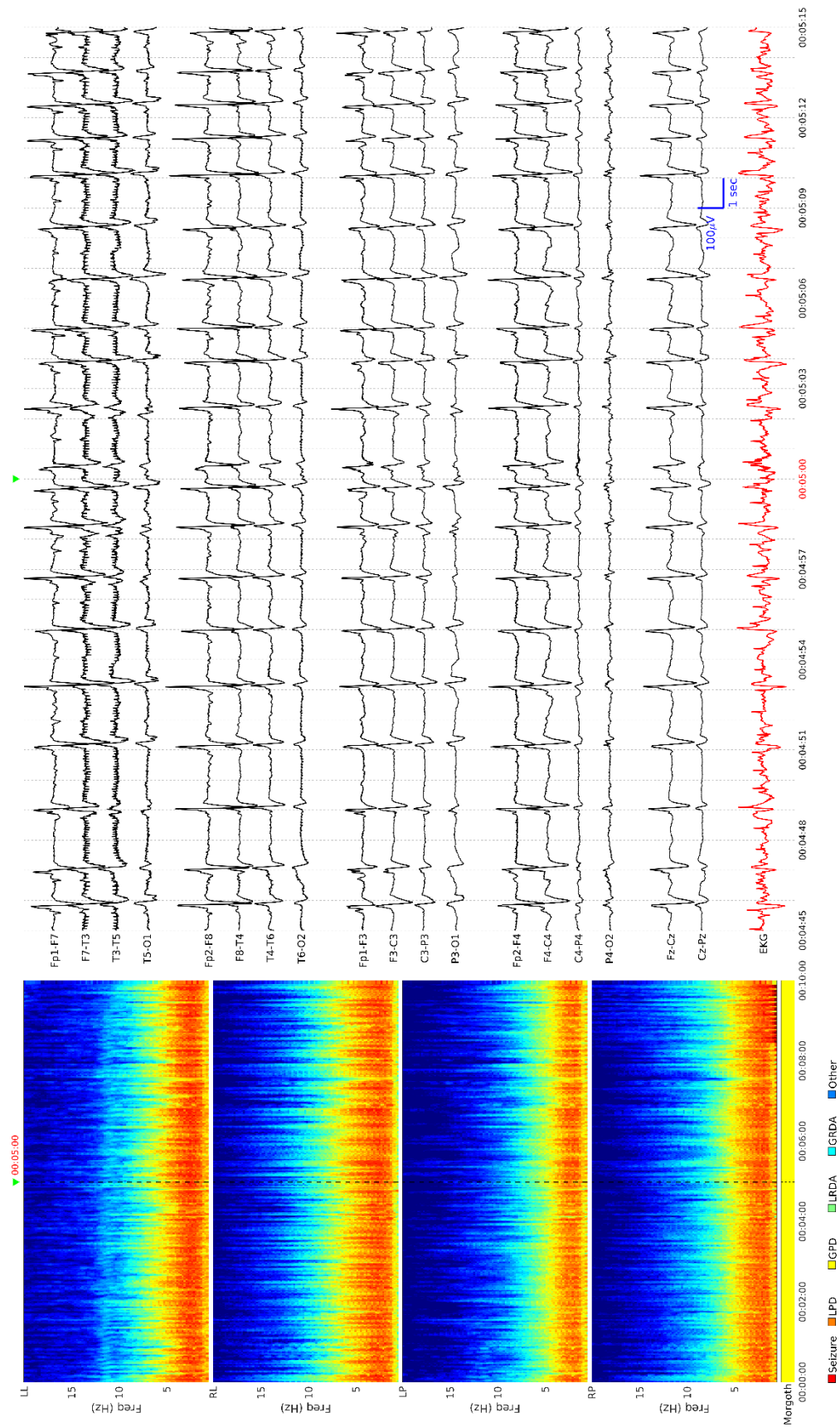


Figure S16. Frontally predominant Generalized Periodic Discharges (GPDs) at 1 Hz on a relatively flat background detected by Morgoth. These manifest in the spectrogram as a generalized broadband monotonous pattern and are indicated in the color bar below the spectrogram in yellow.

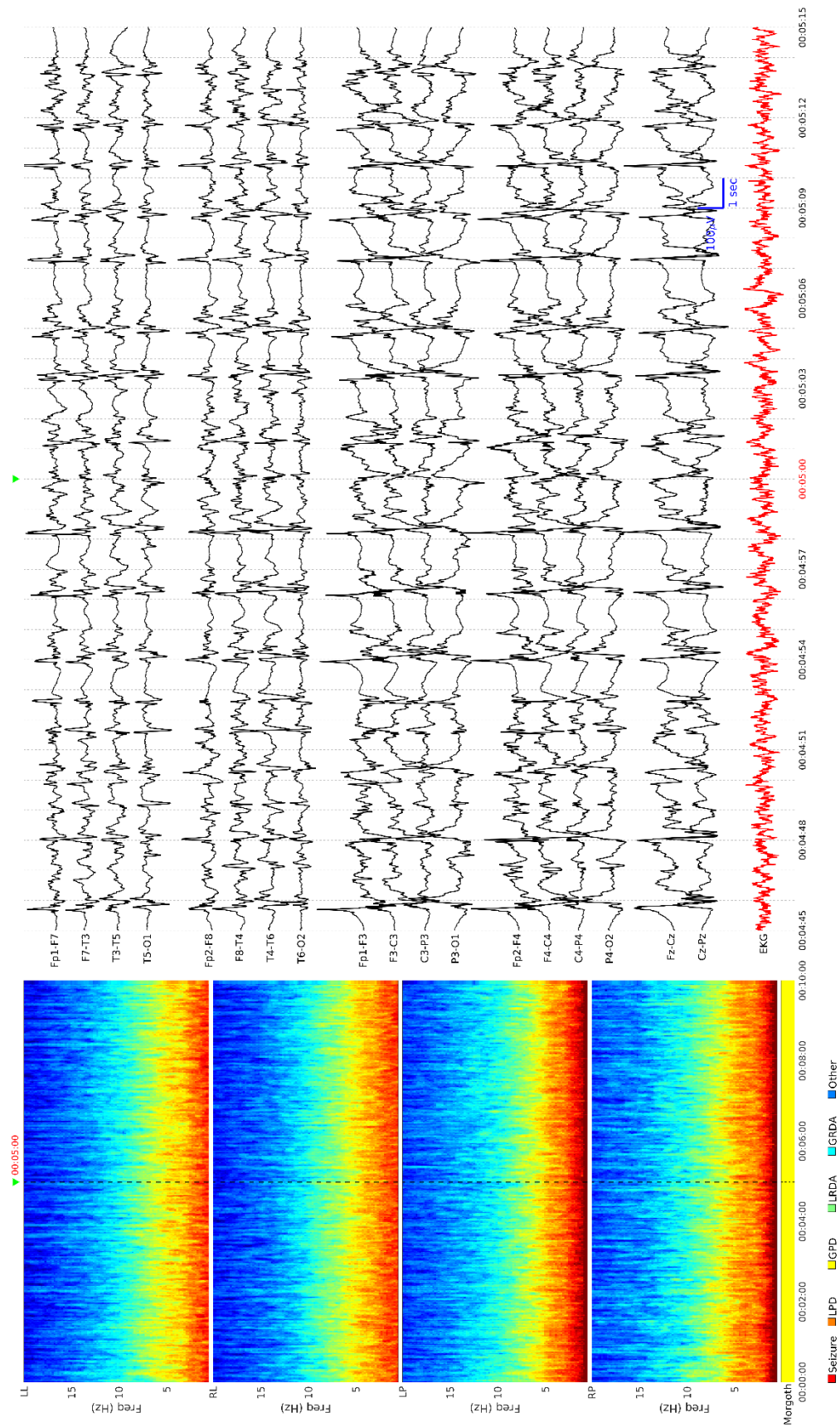


Figure S17. Frontally predominant Generalized Periodic Discharges (GPDs) at approximately 0.5 Hz on a background of generalized irregular delta slowing, detected by Morgoth. These manifest in the spectrogram as a generalized broadband monotonous pattern and are indicated in the color bar below the spectrogram in yellow.

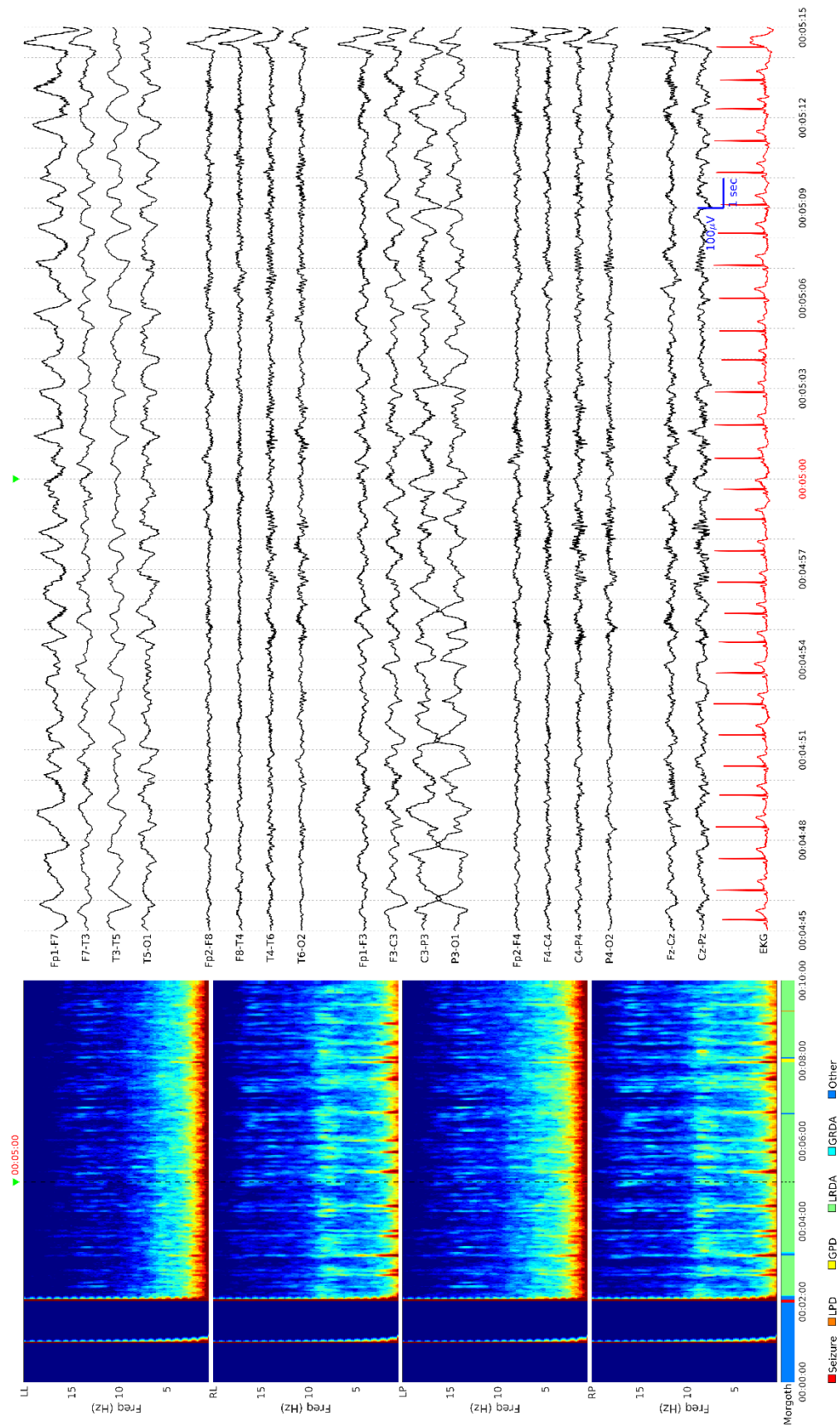


Figure S18. Lateralized Rhythmic Delta Activity (LRDA) detected by Morgho over the left hemisphere. This appears on the spectrogram as increased, relatively narrow-band delta activity, indicated by the green color in the bar below the spectrogram.

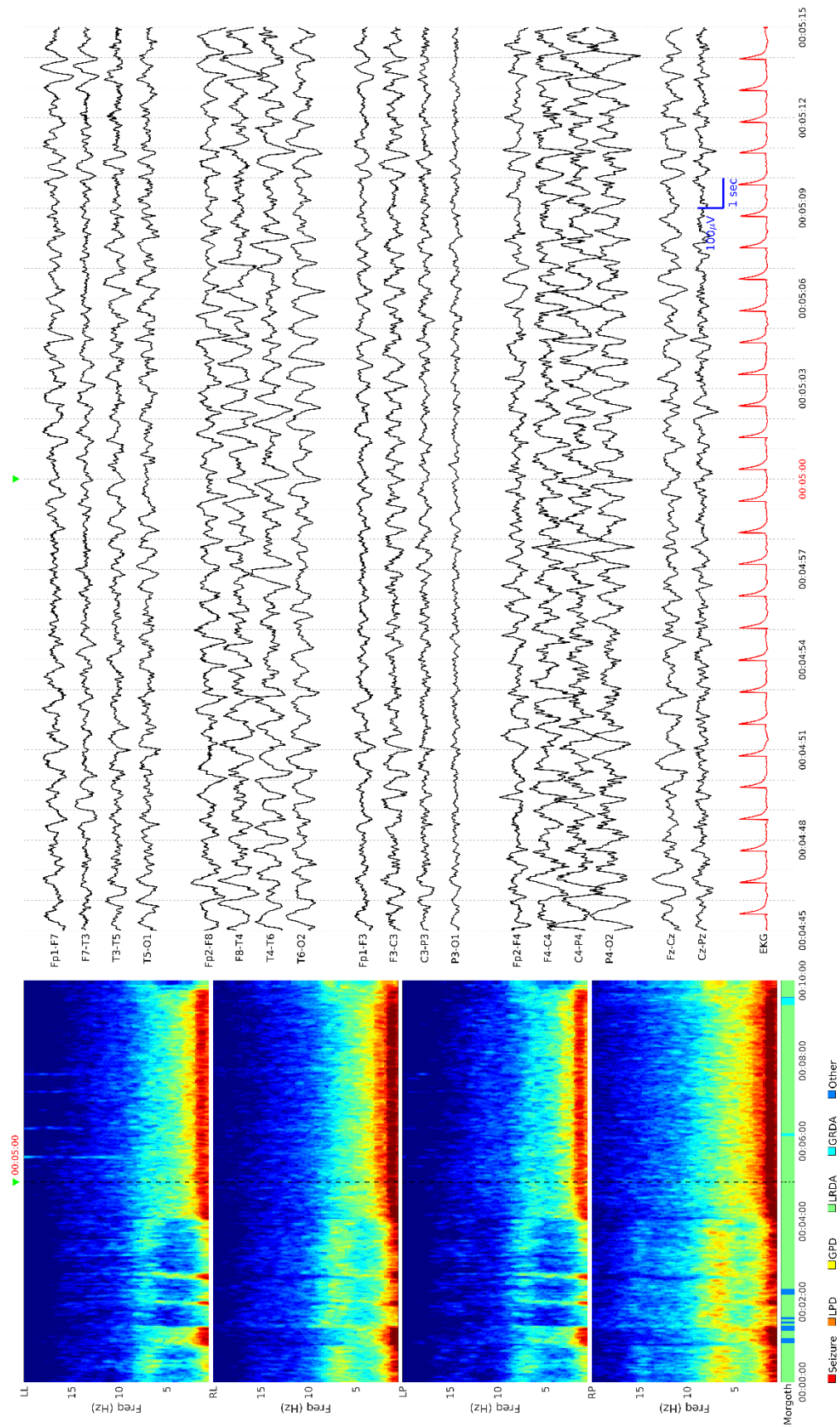


Figure S19. Lateralized Rhythmic Delta Activity (LRDA) detected by Morgoth over the right hemisphere.

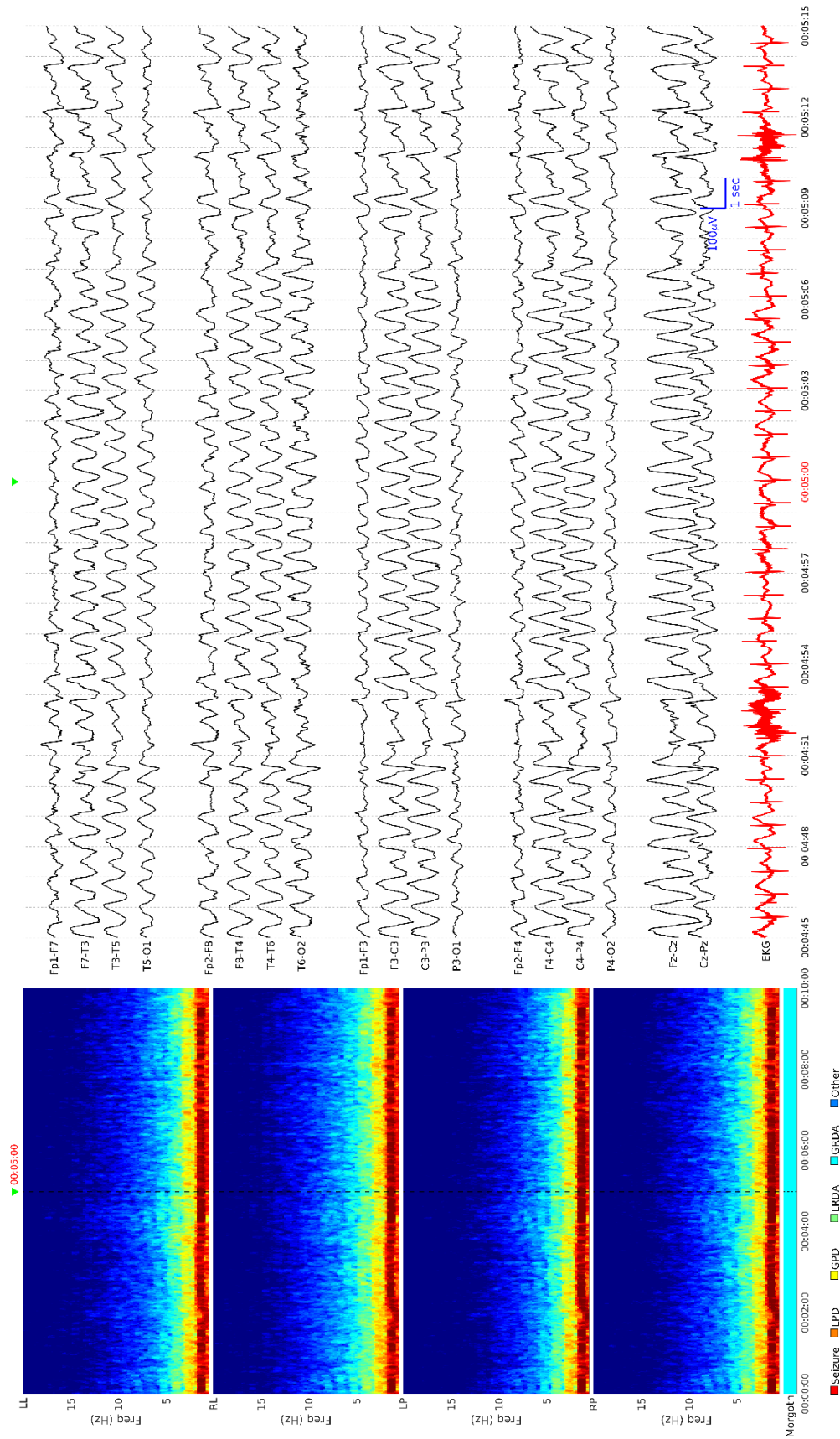


Figure S120. Generalized Rhythmic Delta Activity (GRDA) detected by Morgoth, manifested in the spectrogram as generalized narrow-band delta power. As indicated below the spectrogram by the light blue color bar, in this case GRDA is occurring continuously over the 10 minutes shown.

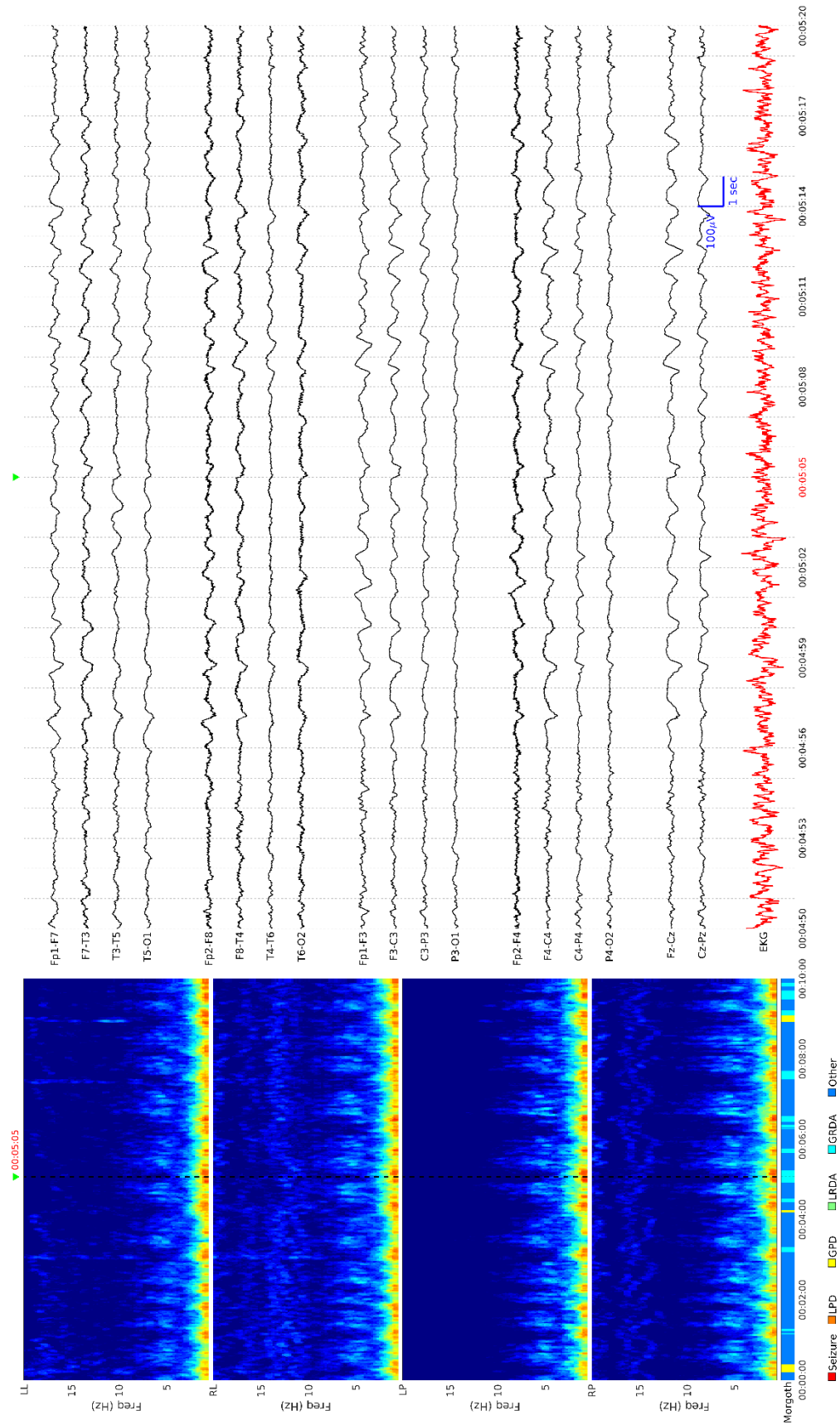


Figure S21. Generalized Rhythmic Delta Activity (GRDA) detected by Morgoth, manifested in the spectrogram as generalized narrow-band delta power. As indicated below the spectrogram by the light blue color bar, in this case GRDA (light blue) is occurring intermittently over the 10 minutes shown; periods in between are not rhythmic or periodic (dark blue).

10. Comprehensive Results

Table S4 lists the comprehensive results of Morgoth, baselines, and experts with pairwise IRRs, leave-one-expert-out IRRs, and ROC, PR for all datasets.

Table S4. Comprehensive Results of Morgoth, Baselines, and Experts: Pairwise IRRs, leave-one-expert-out IRRs, and ROC, PR results for 10 datasets.

(A) Internal HEEDB-Test, MGH-PSG-Test, BCH-PSG-Test datasets. Results are with 95% confidence intervals (CIs), where significance decrease (-) was determined by the absence of overlap between the 95% CIs. HEEDB test set has 14,464 samples with 10,102 Patients; MGH PSG test set has 1,000 Patients; BCH PSG dataset has 2,000 patients.

Test AUC-PR	Train AUC-PR	Test AUC-ROC	Train AUC-ROC
0.981(0.980, 0.982)-	0.999(0.998, 0.999)	0.985(0.984, 0.986) -	1.000(1.000, 1.000)
0.971(0.970, 0.973)-	0.979(0.976, 0.983)	0.984(0.982, 0.986) -	1.000(1.000, 1.000)
0.856(0.848, 0.867)-	0.895(0.886, 0.904)	0.917(0.912, 0.924)	0.923(0.917, 0.929)
0.889(0.880, 0.893)	0.879(0.872, 0.884)	0.896(0.884, 0.905)	0.903(0.891, 0.911)
0.936(0.932, 0.942)	0.950(0.940, 0.960)	0.964(0.955, 0.973)	0.982(0.970, 0.997)
0.928(0.924, 0.934)	0.944(0.930, 0.959)	0.956(0.947, 0.966)	0.975(0.966, 0.983)
0.831(0.817, 0.839)	0.852(0.825, 0.864)	0.964(0.949, 0.971)	0.973(0.969, 0.978)
0.798(0.775, 0.805)	0.808(0.685, 0.995)	0.938(0.915, 0.962)	0.946(0.923, 0.970)
0.845(0.817, 0.878)	0.850(0.822, 0.883)	0.959(0.955, 0.963)	0.964(0.960, 0.968)
0.846(0.818, 0.879)	0.851(0.823, 0.884)	0.959(0.955, 0.963)	0.969(0.959, 0.977)
0.843(0.816, 0.860)	0.848(0.821, 0.865)	0.963(0.951, 0.970)	0.968(0.956, 0.975)
0.809(0.756, 0.878)	0.819(0.766, 0.888)	0.968(0.960, 0.987)	0.978(0.970, 0.997)
0.896(0.888, 0.907)	0.900(0.892, 0.911)	0.968(0.965, 0.973)	0.972(0.969, 0.977)
0.256(0.234, 0.275)	0.261(0.239, 0.285)	0.834(0.821, 0.848)	0.844(0.831, 0.858)
0.909(0.894, 0.926)	0.914(0.899, 0.931)	0.904(0.894, 0.917)	0.909(0.899, 0.922)
0.891(0.883, 0.902)	0.895(0.887, 0.906)	0.963(0.960, 0.968)	0.967(0.964, 0.972)
0.251(0.229, 0.270)	0.256(0.234, 0.280)	0.829(0.816, 0.843)	0.839(0.826, 0.853)
0.904(0.889, 0.921)	0.909(0.894, 0.926)	0.899(0.889, 0.912)	0.904(0.894, 0.917)
0.759(0.728, 0.808)	0.764(0.733, 0.813)	0.964(0.959, 0.967)	0.969(0.964, 0.972)
0.853(0.846, 0.860)	0.858(0.851, 0.865)	0.955(0.954, 0.956)-	0.960(0.959, 0.967)
0.886(0.878, 0.897)	/	0.858(0.855, 0.863)	/
0.146(0.124, 0.165)	/	0.724(0.711, 0.738)	/
0.799(0.784, 0.816)	/	0.794(0.784, 0.807)	/
0.654(0.623, 0.703)	/	0.859(0.854, 0.862)	/
0.748(0.741, 0.755)	/	0.850(0.849, 0.851)	/

EEG-level Class	Test AUC-PR	Train AUC-PR	Test AUC-ROC	Train AUC-ROC	Event-level Class
Normal	0.914(0.912,0.916)-	0.933(0.930,0.936)	0.920(0.913,0.926)-	0.951(0.949,0.953)	Normal
Burst suppression	0.941(0.935,0.947)	0.960(0.945,0.975)	0.969(0.968,0.970)	0.980(0.960,1.000)	Burst suppression
Focal slowing	0.954(0.950,0.957)-	0.966(0.966,0.967)	0.944(0.943,0.946)-	0.985(0.983,0.987)	Focal slowing
Generalized slowing	0.895(0.892,0.898)-	0.922(0.920,0.923)	0.909(0.995,0.923)	0.931(0.920,0.942)	Generalized slowing
Focal spike	0.934(0.919,0.950)	0.967(0.950,0.983)	0.954(0.950,0.958)-	0.974(0.972,0.976)	Focal spike
Generalized spike	0.947(0.941,0.953)-	0.984(0.983,0.984)	0.962(0.951,0.973)-	0.991(0.989,0.992)	Generalized spike
Spikes	0.927(0.900,0.954)	0.942(0.915,0.979))	0.941(0.938,0.954)	0.955(0.940,0.970)	Spikes
Seizure	0.946(0.944,0.947)	0.955(0.943,0.963)	0.954(0.950,0.958)	0.965(0.943,0.987)	Seizure
LPD	0.924(0.914,0.923)-	0.963(0.960,0.966)	0.941(0.937,0.945)-	0.964(0.962,0.966)	LPD
GPD	0.966(0.934,0.997)	0.996(0.995,0.996)	0.981(0.975,0.987)	0.992(0.985,0.999)	GPD
LRDA	0.953(0.935,0.971)	0.981(0.970,0.992)	0.973(0.960,0.986)	0.990(0.994,0.986)	LRDA
GRDA	0.937(0.920,0.954)	0.955(0.954,0.955)	0.965(0.950,0.979)	0.974(0.970,0.978)	GRDA
Awake	0.978(0.977,0.979)	0.999(0.997,0.999)	0.897(0.895,0.900)	0.918(0.915,0.920)	Awake
N1	0.917(0.910,0.922)	0.946(0.940,0.952)	0.564(0.546,0.587)	0.614(0.606,0.627)	N1
N2	0.886(0.879,0.899)	0.919(0.909,0.929)	0.672(0.656,0.689)	0.703(0.686,0.719)	N2
MGH PSG Awake	0.965(0.965,0.970)	0.998(0.995,0.999)	0.718(0.654,0.819)	0.798(0.654,0.899)	MGH PSG Awake
MGH PSG N1	0.966(0.963,0.967)	0.994(0.993,0.997)	0.631(0.561,0.716)	0.643(0.561,0.726)	MGH PSG N1
MGH PSG N2	0.963(0.960,0.966)	0.993(0.990,0.996)	0.676(0.596,0.761)	0.717(0.606,0.801)	MGH PSG N2
MGH PSG N3	0.934(0.930,0.942)	0.965(0.960,0.972)	0.744(0.713,0.774)	0.775(0.743,0.804)	MGH PSG N3
MGH PSG REM	0.533(0.514,0.552)	0.543(0.524,0.562)	0.546(0.534,0.565)	0.567(0.554,0.585)	MGH PSG REM
BCH PSG Awake	0.967(0.967,0.969)	0.999(0.997,0.999)	0.762(0.748,0.776)	0.773(0.758,0.786)	BCH PSG Awake
BCH PSG N1	0.945(0.934,0.953)	0.976(0.964,0.983)	0.803(0.768,0.840)	0.843(0.808,0.880)	BCH PSG N1
BCH PSG N2	0.939(0.931,0.947)	0.970(0.961,0.977)	0.656(0.620,0.696)	0.688(0.650,0.726)	BCH PSG N2
BCH PSG N3	0.938(0.932,0.945)	0.969(0.962,0.975)	0.800(0.778,0.834)	0.829(0.798,0.854)	BCH PSG N3
BCH PSG REM	0.933(0.906,0.954)	0.974(0.966,0.984)	0.542(0.445,0.612)	0.554(0.455,0.622)	BCH PSG REM

(B) Internal multi-expert scored datasets MoE. Comparisons were based on the 95% confidence intervals (CIs), where significance (*) was determined by the absence of overlap between the 95% CIs. Experts have labeled at least 1,000 samples (2,761 samples with 2,250 patients in total). Pairwise IRR analysis is limited to experts who labeled at least 100 samples in common per class. Based on this, MoE dataset has 21 experts with 394 expert pairs.

Morgoth AUC-ROC	Average agreement between Morgoth and the consensus with one expert left out	Average agreement between one expert and the consensus with the expert left out	Average agreement between Morgoth with one expert	Average agreement across all pairwise expert comparisons ¹
0.992(0.989, 0.994)	0.797(0.770, 0.824) *	0.703(0.636, 0.769)	0.688(0.622, 0.755)	0.589(0.494, 0.684)
0.920(0.897, 0.935)	0.986(0.977, 0.996) *	0.547(0.474, 0.620)	0.516(0.436, 0.596)	0.416(0.325, 0.508)
0.899(0.887, 0.910)	0.845(0.828, 0.862) *	0.438(0.387, 0.488)	0.414(0.363, 0.465) *	0.267(0.202, 0.332)
0.967(0.954, 0.977)	0.932(0.904, 0.961) *	0.628(0.538, 0.717)	0.583(0.487, 0.680)	0.542(0.435, 0.649)
0.972(0.962, 0.981)	0.839(0.793, 0.885) *	0.593(0.489, 0.696)	0.552(0.440, 0.665)	0.514(0.385, 0.643)
0.956(0.922, 0.982)	0.888(0.806, 0.970) *	0.428(0.188, 0.669)	0.447(0.203, 0.690)	0.362(0.089, 0.636)
0.970(0.946, 0.988)	0.920(0.882, 0.958) *	0.562(0.408, 0.716)	0.581(0.438, 0.724)	0.483(0.298, 0.669)
0.972(0.956, 0.986)	0.941(0.907, 0.975) *	0.579(0.437, 0.721)	0.558(0.404, 0.711)	0.472(0.296, 0.648)
0.982(0.953, 0.999)	0.941(0.894, 0.999) *	0.379(0.077, 0.680)	0.396(0.097, 0.694)	0.286(-0.043, 0.615)
0.984(0.965, 0.998)	0.968(0.914, 1.021) *	0.300(-0.071, 0.671)	0.327(-0.064, 0.718)	0.205(-0.197, 0.607)
0.987(0.979, 0.994)	0.832(0.786, 0.879) *	0.671(0.558, 0.784)	0.619(0.526, 0.712)	0.515(0.357, 0.673)
0.699(0.640, 0.773)	0.003(-0.235, 0.240)	0.499(0.219, 0.779)	0.021(-0.277, 0.318)	0.305(-0.030, 0.641)
0.983(0.973, 0.991)	0.735(0.675, 0.794)	0.657(0.521, 0.792)	0.497(0.386, 0.608)	0.493(0.287, 0.698)

Class	Morgoth EUC-PR	Morgoth AUC-PR	Morgoth EUC-ROC
Burst suppression	100.0%(100.0%, 100.0%)	0.956(0.944, 0.966)	97.7%(90.5%, 100.0%)
Focal slowing	90.9%(85.7%, 95.2%)	0.820(0.780, 0.851)	89.8%(85.7%, 95.2%)
Generalized slowing	100.0%(100.0%, 100.0%)	0.884(0.869, 0.898)	99.1%(95.2%, 100.0%)
Focal spike	90.0%(81.0%, 95.2%)	0.847(0.803, 0.888)	88.6%(76.2%, 95.2%)
Generalized spike	97.0%(90.5%, 100.0%)	0.827(0.775, 0.869)	95.4%(90.5%, 100.0%)
Seizure	91.3%(81.0%, 100.0%)	0.722(0.566, 0.834)	46.9%(28.6%, 71.4%)
LPD	99.2%(95.2%, 100.0%)	0.857(0.797, 0.912)	74.5%(61.9%, 90.5%)
GPD	97.1%(85.7%, 100.0%)	0.852(0.782, 0.897)	85.0%(66.7%, 95.2%)
LRDA	97.7%(90.5%, 100.0%)	0.836(0.679, 0.951)	68.2%(52.4%, 81.0%)
GRDA	97.7%(85.7%, 100.0%)	0.793(0.574, 0.944)	60.0%(38.1%, 76.2%)
Awake	96.1%(90.5%, 100.0%)	0.995(0.992, 0.998)	96.5%(90.5%, 100.0%)
Sleep N1	8.3%(4.8%, 14.3%)	0.131(0.080, 0.220)	8.1%(4.8%, 14.3%)
Sleep N2	69.1%(38.1%, 90.5%)	0.919(0.871, 0.958)	59.9%(35.6%, 85.7%)

(C) Internal multi-expert scored datasets IIIC-Test. Comparisons were based on the 95% confidence intervals (CIs), where significance (*) was determined by the absence of overlap between the 95% CIs. Experts have labeled at least 1,000 samples (15,747 samples with 758 patients in total). Based on this, IIIC-Test dataset has 30 experts with 772 expert pairs.

Average agreement across all pairwise expert comparisons [†]
0.368(0.205, 0.531)
0.496(0.419, 0.574)
0.450(0.338, 0.563)
0.253(0.094, 0.412)
0.345(0.221, 0.468)
0.367(0.305, 0.430)

Average agreement between Kaggle 1st and the consensus with one expert left out	Average agreement between SPaRCNet and the consensus with one expert left out	Average agreement between one expert and the consensus with the expert left out	Average agreement between Morgoth with one expert	Average agreement between Kaggle 1st with one expert	Average agreement between SPaRCNet with one expert
0.698(0.684, 0.713)	0.422(0.408, 0.436)	0.496(0.398, 0.594)	0.460(0.364, 0.557)	0.433(0.325, 0.541)	0.220(0.130, 0.309)
0.569(0.554, 0.584)	0.480(0.463, 0.496)	0.617(0.574, 0.660)	0.627(0.585, 0.669) *	0.566(0.519, 0.612)	0.447(0.394, 0.500)
0.675(0.660, 0.702)*	0.591(0.566, 0.616)	0.596(0.533, 0.659)	0.597(0.535, 0.659)	0.572(0.510, 0.634)	0.470(0.398, 0.543)
0.433(0.395, 0.470)	0.301(0.264, 0.338)	0.336(0.230, 0.442)	0.382(0.280, 0.483)	0.321(0.213, 0.428)	0.270(0.171, 0.369)
0.592(0.569, 0.615)*	0.433(0.410, 0.456)	0.452(0.376, 0.528)	0.485(0.412, 0.558)	0.440(0.366, 0.515)	0.359(0.294, 0.424)
0.479(0.467, 0.490)	0.354(0.341, 0.367)	0.506(0.470, 0.543)	0.518(0.482, 0.554) *	0.447(0.410, 0.484)	0.287(0.246, 0.328)

Kaggle 1 st EUC-ROC	SPaRCNet EUC-ROC	Morgoth AUC-ROC	Kaggle 1 st AUC-ROC	SPaRCNet AUC-ROC	Average agreement between Morgoth and the consensus with one expert left out
29.4%(20.0%, 40.0%)	0.0%(0.0%, 0.0%)	0.988(0.986, 0.989)*	0.955(0.951, 0.960)	0.856(0.847, 0.863)	0.728(0.722, 0.733)
35.5%(26.7%, 43.3%)	0.0%(0.0%, 0.0%)	0.990(0.988, 0.991)*	0.947(0.942, 0.951)	0.851(0.843, 0.858)	0.917(0.910, 0.924)*
97.4%(91.6%, 100.0%)	40.3%(26.7%, 55.1%)	0.994(0.992, 0.995)*	0.979(0.975, 0.983)	0.932(0.923, 0.939)	0.899(0.886, 0.911)*
69.7%(54.9%, 83.3%)	21.6%(10.0%, 35.1%)	0.991(0.986, 0.994)*	0.939(0.929, 0.948)	0.866(0.848, 0.883)	0.846(0.826, 0.866)*
69.3%(58.2%, 76.7%)	14.7%(8.2%, 23.3%)	0.990(0.988, 0.993)*	0.965(0.961, 0.968)	0.901(0.890, 0.912)	0.859(0.845, 0.873)*
53.1%(46.7%, 60.0%)	0.0%(0.0%, 0.0%)	0.987(0.985, 0.989)*	0.907(0.902, 0.913)	0.780(0.770, 0.790)	0.890(0.884, 0.897)*

Kaggle 1 st EUC-PR	SPaRCNet EUC-PR	Morgoth AUC-PR	Kaggle 1 st AUC-PR	SPaRCNet AUC-PR	Morgoth EUC-ROC
69.3% (60.0%, 76.7%)	11.1% (6.7%, 16.7%)	0.868 (0.837, 0.900)*	0.721 (0.687, 0.750)	0.305 (0.281, 0.331)	82.1%(73.3%, 90.0%)*
49.6% (36.7%, 60.0%)	0.0% (0.0%, 0.0%)	0.980 (0.976, 0.983)*	0.870 (0.860, 0.882)	0.707 (0.688, 0.725)	100.0%(100.0%, 100.0%)*
85.3% (73.3%, 93.3%)	30.0% (16.7%, 40.0%)	0.960 (0.952, 0.967)*	0.875 (0.858, 0.891)	0.739 (0.708, 0.765)	100.0%(100.0%, 100.0%)*
67.1% (53.3%, 80.0%)	25.6% (13.3%, 38.4%)	0.911 (0.887, 0.932)*	0.556 (0.507, 0.601)	0.377 (0.331, 0.423)	100.0%(100.0%, 100.0%)*
65.5% (54.9%, 78.4%)	12.2% (4.9%, 20.0%)	0.937 (0.924, 0.950)*	0.728 (0.700, 0.756)	0.524 (0.485, 0.560)	97.7%(96.7%, 100.0%)*
40.5% (31.6%, 50.0%)	0.1% (0.0%, 0.0%)	0.982 (0.979, 0.985)*	0.882 (0.874, 0.890)	0.668 (0.653, 0.684)	100.0%(100.0%, 100.0%)*

Class	Morgoth EUC-PR
Seizure	96.2% (90.0%, 100.0%)*
LPD	100.0% (100.0%, 100.0%)*
GPD	99.9% (98.2%, 100.0%)*
LRDA	100.0% (100.0%, 100.0%)*
GRDA	97.9% (96.7%, 100.0%)*
Other	100.0% (100.0%, 100.0%)*

(D) Internal multi-expert scored datasets SN2-Test. Comparisons were based on the 95% confidence intervals (CIs), where significance (*) was determined by the absence of overlap between the 95% CIs. Experts have labeled at least 1,000 samples (with 1,602 samples with 208 patients in total). Pairwise IRR analysis is limited to experts who labeled at least 100 samples in common per class. Based on this, SN2-Test dataset has 8 experts with 56 expert pairs.

Average agreement between Morgoth with one expert	Average agreement between SpikeNet 2 with one expert	Average agreement across all pairwise expert comparisons ¹
0.567(0.479, 0.656)*	0.501(0.400, 0.601)	0.370(0.320, 0.419)

SpikeNet 2 EUC-ROC	Morgoth AUC-ROC	SpikeNet 2 AUC-ROC	Average agreement between Morgoth and the consensus with one expert left out	Average agreement between SpikeNet 2 and the consensus with one expert left out	Average agreement between one expert and the consensus with the expert left out
73.3%(54.2%, 87.5%)	1.000(1.000, 1.000)*	0.974(0.963, 0.985)	0.988(0.981, 0.996))*	0.924(0.873, 0.975))*	0.572(0.484, 0.659)

Class	Morgoth EUC-PR	SpikeNet2 EUC-PR	Morgoth AUC-PR	SpikeNet2 AUC-PR	Morgoth EUC-ROC
Spikes	99.3%(95.8%, 100.0%)	82.4%(66.7%, 95.8%)	1.000(1.000, 1.000)*	0.990(0.985, 0.995)	99.9%(95.8%, 100.0%)*

(E) External single-score datasets TUH. Results are with 95% confidence intervals (CIs), where significance (*) was determined by the absence of overlap between the 95% CIs.

Morgoth AUC-ROC	LaBraM AUC-ROC	EEGformer AUC-ROC	BrainBERT AUC-ROC	BIOT AUC-ROC	GraphS4mer AUC-ROC
0.902(0.900, 0.904) *	0.866(0.864, 0.868)	0.882(0.880, 0.884)	0.875(0.873, 0.877)	0.862(0.860, 0.864)	0.860(0.858, 0.862)
0.883(0.881, 0.885)	0.893(0.891, 0.895)	0.883(0.881, 0.885)	0.835(0.833, 0.837)	0.844(0.842, 0.846)	0.826(0.824, 0.828)
0.741(0.739, 0.743)	0.731(0.723, 0.739)	0.722(0.720, 0.724)	0.701(0.699, 0.703)	0.683(0.681, 0.685)	0.636(0.634, 0.638)
0.705(0.703, 0.707)	0.704(0.702, 0.706)	0.703(0.701, 0.705)	0.685(0.683, 0.687)	0.653(0.651, 0.655)	0.666(0.664, 0.668)

776
777

(F) External single-score datasets MASS.

Task	Class	Morgoth AUC-ROC
TUAB	Normal	0.902(0.900, 0.904)
TUSZ	Seizure	0.883(0.881, 0.885)
TUSL	Seizure	0.747(0.715, 0.779)
	Slowing	0.729(0.695, 0.763)
	Background	0.758(0.732, 0.784)
TUEV	Spike and slow wave	0.685(0.671, 0.699)
	GPD	0.712(0.690, 0.734)
	LPD	0.685(0.671, 0.699)
	Eye movement	0.678(0.652, 0.704)
	Artifact	0.721(0.702, 0.740)
	Background	0.693(0.667, 0.719)

U-Sleep AUC-PR	Morgoth AUC-ROC	U-Sleep AUC-ROC
0.931 (0.928, 0.933)	0.985 (0.984, 0.986)	0.973 (0.971, 0.974)
0.341 (0.335, 0.348)	0.845 (0.842, 0.848)	0.807 (0.803, 0.810)
0.867 (0.865, 0.869)	0.917 (0.916, 0.919)	0.896 (0.895, 0.898)
0.770 (0.765, 0.776)	0.972 (0.972, 0.974)	0.960 (0.958, 0.961)
0.825 (0.820, 0.829)	0.977 (0.976, 0.978)	0.923 (0.920, 0.925)

Class	Morgoth AUC-PR
Awake	0.949 (0.947, 0.951)
N1	0.374 (0.368, 0.381)
N2	0.891 (0.889, 0.893)
N3	0.839 (0.835, 0.844)
REM	0.888 (0.885, 0.892)

(G) External multi-expert scored datasets HEP. Comparisons were based on the 95% confidence intervals (CIs), where significance (*) was determined by the absence of overlap between the 95% CIs.

Morgoth AUC-PR	SpikeNet2 AUC-PR	Morgoth AUC-ROC	SpikeNet2 AUC-ROC
0.999(0.999, 1.000)	0.945(0.938, 0.952)	0.999 (0.998, 1.000)	0.941 (0.933, 0.949)

Task	Morgoth AUC-PR	Morgoth AUC-ROC	Task
Event-level Spikes	0.887 (0.819, 0.941)	0.927 (0.893, 0.960)	EEG-level Spikes

(H) External multi-expert scored datasets SAI. *Comparisons were based on the 95% confidence intervals (CIs), where significance (*) was determined by the absence of overlap between the 95% CIs. Experts have labeled at least 30 patients (100 patients in total). Pairwise IRR analysis is limited to experts who labeled at least 10 patients in common per class. Based on this, SCORE-AI test dataset has 14 experts with 164 expert pairs.

Average agreement across all pairwise expert comparisons ¹
0.724(0.544, 0.903)
0.484(0.226, 0.743)
0.461(0.193, 0.729)
0.531(0.275, 0.787)
0.686(0.422, 0.949)

Average agreement between Morgoth and the consensus with one expert left out	Average agreement between SCORE-AI and the consensus with one expert left out	Average agreement between one expert and the consensus with the expert left out	Average agreement between Morgoth with one expert	Average agreement between SCORE-AI one expert
0.897(0.822, 0.971)	0.899(0.827, 0.972)	0.823(0.695, 0.950)	0.736(0.579, 0.892)	0.779(0.636, 0.921)
0.260(0.050, 0.471)	0.010(-0.217, 0.237)	0.602(0.394, 0.810)	0.538(0.295, 0.780)	0.456(0.211, 0.701)
0.665(0.508, 0.822)	0.616(0.471, 0.761)	0.542(0.316, 0.768)	0.464(0.182, 0.746)	0.489(0.251, 0.727)
0.600(0.422, 0.778)	0.595(0.436, 0.754)	0.630(0.416, 0.843)	0.459(0.185, 0.733)	0.493(0.256, 0.731)
0.751(0.589, 0.913)	0.828(0.704, 0.951)	0.789(0.594, 0.984)	0.650(0.384, 0.917)	0.739(0.534, 0.943)

Morgoth AUC-PR	SCORE-AI AUC-PR	Morgoth EUC-ROC	SCORE-AI EUC-ROC	Morgoth AUC-ROC	SCORE-AI AUC-ROC
0.973(0.884, 1.000)	0.994(0.968, 1.000)	51.8%(21.4%, 78.6%)	57.4%(42.9%, 78.6%)	0.987(0.950, 1.000)	0.995(0.971, 1.000)
0.958(0.703, 1.000)	0.769(0.507, 0.962)	72.3%(57.1%, 85.7%)	28.5%(3.4%, 57.1%)	0.978(0.890, 1.000)	0.869(0.760, 0.978)
0.902(0.753, 1.000)	0.798(0.576, 0.960)	45.7%(24.6%, 57.1%)	31.0%(7.1%, 57.1%)	0.951(0.880, 1.000)	0.928(0.859, 0.985)
0.878(0.687, 1.000)	0.810(0.527, 1.000)	56.6%(14.3%, 92.9%)	43.3%(3.4%, 78.6%)	0.945(0.852, 1.000)	0.929(0.846, 1.000)
0.879(0.666, 1.000)	0.965(0.810, 1.000)	26.7%(3.4%, 71.4%)	36.9%(7.1%, 78.6%)	0.983(0.933, 1.000)	0.993(0.971, 1.000)

Class	Morgoth EUC-PR	SCORE-AI EUC-PR
Normal	63.2%(28.6%, 92.9%)	69.1%(42.9%, 92.9%)
Focal Slowing	71.7%(50.0%, 85.7%)	29.7%(7.1%, 57.1%)
Generalized Slowing	48.6%(31.8%, 62.7%)	28.7%(10.5%, 50.0%)
Focal Spike	58.1%(21.4%, 89.5%)	43.9%(7.1%, 82.3%)
Generalized Spike	25.9%(0.0%, 57.1%)	36.4%(14.3%, 57.1%)

- (I) External multi-expert scored datasets ON. Comparisons were based on the 95% confidence intervals (CIs), where significance (*) was determined by the absence of overlap between the 95% CIs. Experts have labeled at least 30 patients (100 patients in total). Pairwise IRR analysis is limited to experts who labeled at least 10 patients in common per class. Based on this, ON dataset has 14 experts with 182 expert pairs.

Average agreement between Morgoth with one expert	Average agreement across all pairwise expert comparisons ¹
0.281(-0.019, 0.581)	0.459(0.238, 0.68)
0.378(0.073, 0.682)	0.391(0.14, 0.642)
0.619(0.418, 0.820)	0.646(0.453, 0.84))
0.807(0.659, 0.955)	0.805(0.653, 0.958)

Morgoth EUC-PR	Morgoth AUC-PR	Morgoth EUC-ROC	Morgoth AUC-ROC	Average agreement between Morgoth and the consensus with one expert left out	Average agreement between one expert and the consensus with the expert left out
54.1%(14.3%, 85.7%)	0.833(0.639, 0.971)	60.7%(21.4%, 92.9%)	0.925(0.820, 0.995)	0.449(0.205, 0.694)	0.582(0.380, 0.783)
31.3%(0.0%, 78.6%)	0.768(0.597, 0.901)	30.5%(0.0%, 78.6%)	0.901(0.827, 0.962)	0.562(0.319, 0.806)	0.559(0.328, 0.790)
24.8%(0.0%, 64.3%)	0.854(0.701, 0.960)	23.1%(0.0%, 64.3%)	0.935(0.865, 0.985)	0.739(0.594, 0.884)	0.735(0.564, 0.907)
45.8%(21.4%, 71.4%)	0.982(0.941, 1.000)	61.0%(35.7%, 78.6%)	0.997(0.989, 1.000)	0.834(0.715, 0.953)	0.837(0.696, 0.978)

Class
Focal slowing
Generalized slowing
Focal spike
Generalized spike

(J) External multi-expert scored datasets UPenn. *Comparisons were based on the 95% confidence intervals (CIs), where significance (*) was determined by the absence of overlap between the 95% CIs. Experts have labeled 67 patients (67 patients in total). Based on this, UPenn dataset has 6 experts with 30 expert pairs.

Average agreement between Morgoth with one expert	Average agreement between U-Sleep2 with one expert	Average agreement between U-Sleep1 with one expert	Average agreement across all pairwise expert comparisons ¹
0.733(0.722, 0.745)*	0.711(0.699, 0.722)	0.710(0.698, 0.721)	0.701(0.689, 0.712)
0.160(0.128, 0.193)	0.148(0.117, 0.179)	0.154(0.123, 0.185)	0.188(0.163, 0.213)
0.599(0.590, 0.608)	0.560(0.551, 0.569)	0.562(0.553, 0.571)	0.528(0.489, 0.597)
0.467(0.446, 0.488)*	0.431(0.408, 0.453)	0.438(0.417, 0.460)	0.406(0.387, 0.424)
0.741(0.731, 0.751)	0.682(0.671, 0.693)	0.685(0.674, 0.696)	0.764(0.755, 0.774)

U-Sleep2 AUC-ROC	U-Sleep1 AUC-ROC	Average agreement between Morgoth and the consensus with one expert left out	Average agreement between U-Sleep2 and the consensus with one expert left out	Average agreement between U-Sleep1 and the consensus with one expert left out	Average agreement between one expert and the consensus with the expert left out
0.950(0.946, 0.953)	0.951(0.949, 0.954)	0.803(0.795, 0.811)*	0.782(0.773, 0.791)	0.785(0.776, 0.793)	0.776(0.766, 0.786)
0.808(0.798, 0.814)	0.808(0.799, 0.815)	0.317(0.249, 0.385)	0.287(0.260, 0.314)	0.295(0.268, 0.322)*	0.234(0.206, 0.262)
0.901(0.899, 0.904)	0.899(0.897, 0.903)	0.716(0.710, 0.723)*	0.674(0.667, 0.681)*	0.675(0.668, 0.681)*	0.642(0.634, 0.651)
0.967(0.965, 0.969)	0.968(0.966, 0.969)	0.607(0.591, 0.623)*	0.561(0.544, 0.577)*	0.562(0.545, 0.578)*	0.494(0.476, 0.513)
0.958(0.956, 0.961)	0.959(0.958, 0.962)	0.809(0.802, 0.816)	0.751(0.743, 0.759)	0.753(0.744, 0.761)	0.827(0.818, 0.835)

Morgoth AUC-PR	U-Sleep2 AUC-PR	U-Sleep1 AUC-PR	Morgoth EUC-ROC	U-Sleep2 EUC-ROC	U-Sleep1 EUC-ROC	Morgoth AUC-ROC
0.890(0.885, 0.896)	0.871(0.865, 0.877)	0.873(0.868, 0.879)	50.0%(50.0%, 50.0%)	47.3%(33.3%, 50.0%)	50.0%(50.0%, 50.0%)	0.962(0.960, 0.965)
0.262(0.248, 0.278)	0.201(0.189, 0.213)	0.201(0.191, 0.211)	58.7%(33.3%, 83.3%)	20.7%(16.7%, 33.3%)	20.0%(16.7%, 33.3%)	0.839(0.830, 0.845)
0.914(0.912, 0.918)	0.905(0.902, 0.909)	0.903(0.900, 0.907)	33.7%(33.3%, 33.3%)	33.3%(33.3%, 33.3%)	33.3%(33.3%, 33.3%)	0.910(0.907, 0.913)
0.770(0.758, 0.782)	0.705(0.691, 0.722)	0.711(0.696, 0.727)	66.7%(66.7%, 66.7%)	51.7%(50.0%, 66.7%)	55.0%(50.0%, 66.7%)	0.974(0.973, 0.975)
0.857(0.849, 0.864)	0.838(0.829, 0.845)	0.843(0.834, 0.850)	0.0%(0.0%, 0.0%)	0.0%(0.0%, 0.0%)	0.0%(0.0%, 0.0%)	0.965(0.963, 0.967)

Class	Morgoth EUC-PR	U-Sleep2 EUC-PR	U-Sleep1 EUC-PR
Awake	50.0%(50.0%, 50.0%)	48.7%(33.3%, 50.0%)	50.0%(50.0%, 50.0%)
N1	52.7%(50.0%, 66.7%)	33.3%(16.7%, 50.0%)	33.7%(16.7%, 50.0%)
N2	33.3%(33.3%, 33.3%)	33.3%(33.3%, 33.3%)	33.3%(33.3%, 33.3%)
N3	82.0%(66.7%, 83.3%)	52.0%(50.0%, 66.7%)	55.3%(50.0%, 66.7%)
REM	0.0%(0.0%, 0.0%)	0.0%(0.0%, 0.0%)	0.0%(0.0%, 0.0%)

11. Subgroup Results

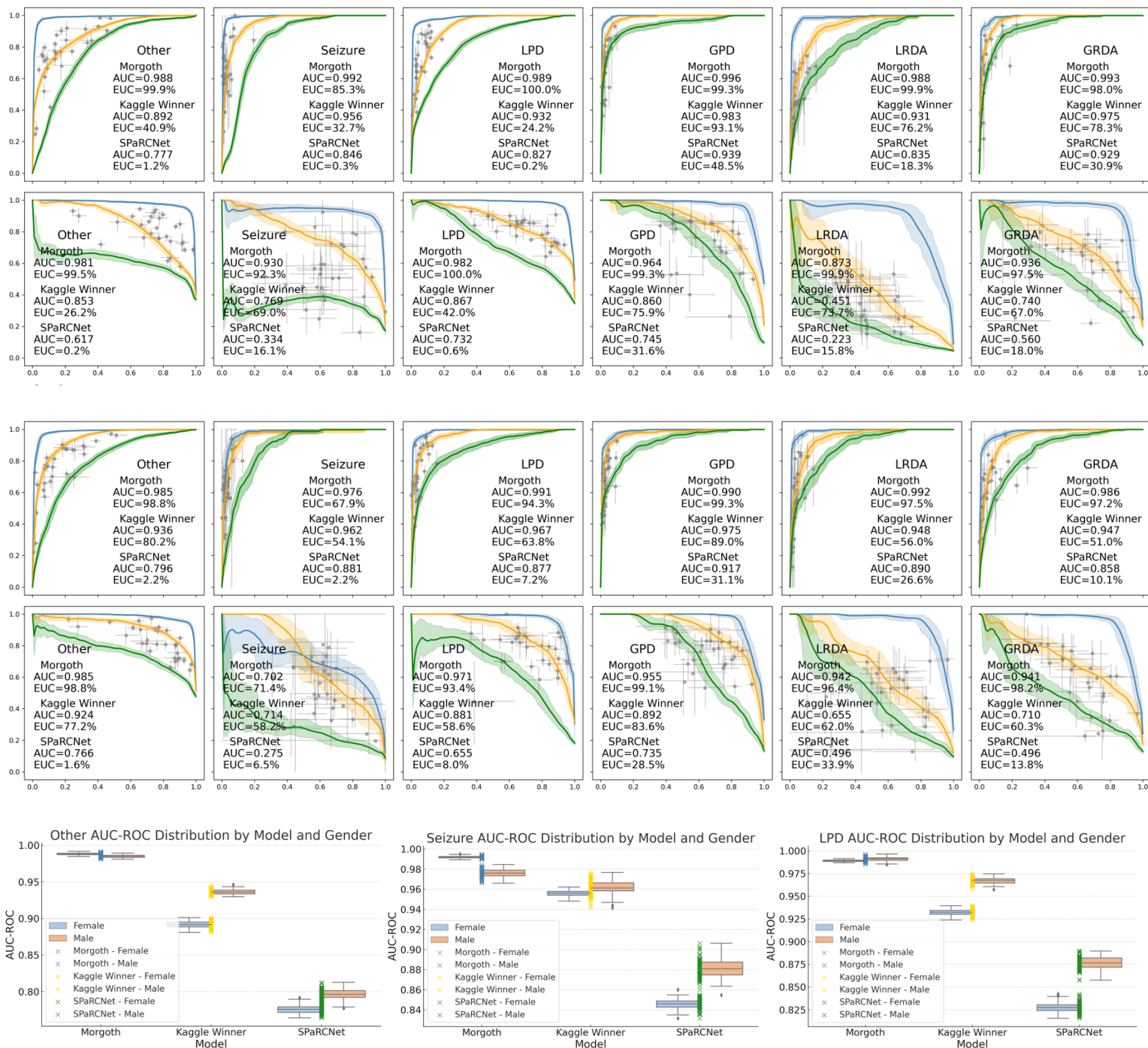
11.1 Effects of on sex on algorithm performance

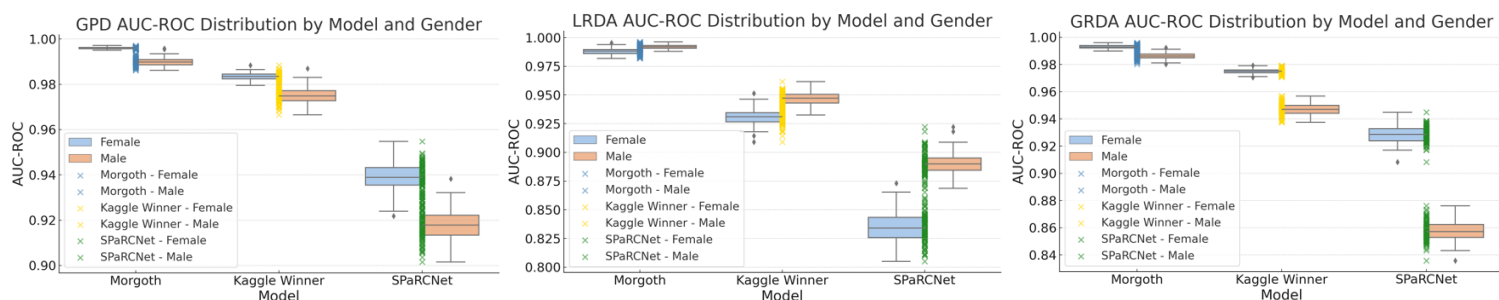
We performed gender subgroup analyses on the multi-expert scored datasets: IIIC-Test, SN2-Test, MoE, SAI, and ON datasets (the UPenn dataset was excluded from subgroup analysis as it consisted entirely of female subjects). Overall results are shown in Figure S22 with female results on the left/top and male results on the right/bottom.

All detailed results are shown in Table S4. All significant test results are shown in Table S5. MORGOTH showed smaller differences between gender groups compared to the baseline models, demonstrating good generalization and robustness

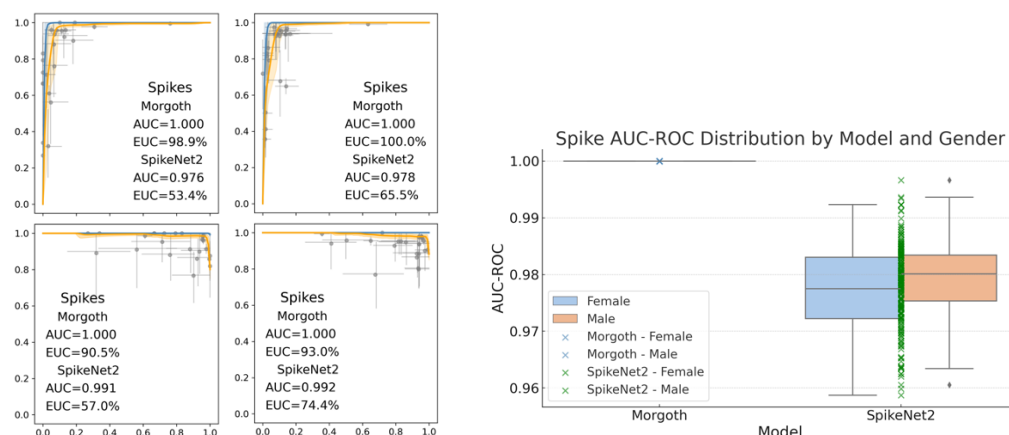
For the IIIC-Test dataset, the proportion of cases showing significant differences ($P < 0.05$) was 33.3% for Morgoth, 44% for SPaRCNet, and 48% for the Kaggle winner; For the SN2-Test dataset, both MORGOTH and SPaRCNet2 showed 0% of cases with significant differences, and MORGOTH generally had higher p-values; For MoE dataset, MORGOTH did not show significant differences across groups; For SAI dataset, both MORGOTH and SCORE-AI showed

817 0% of cases with significant differences; For the SAI dataset, both MORGOTH and SCORE-AI
818 showed 0% of cases with significant differences; For the ON dataset, MORGOTH has 25%
819 differences.

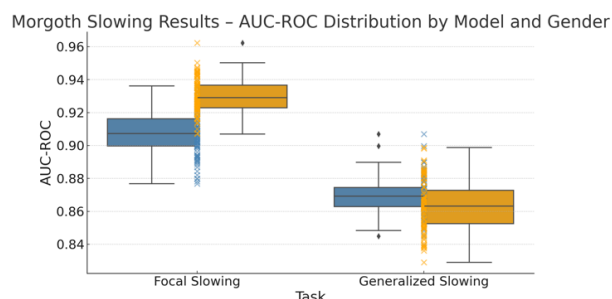
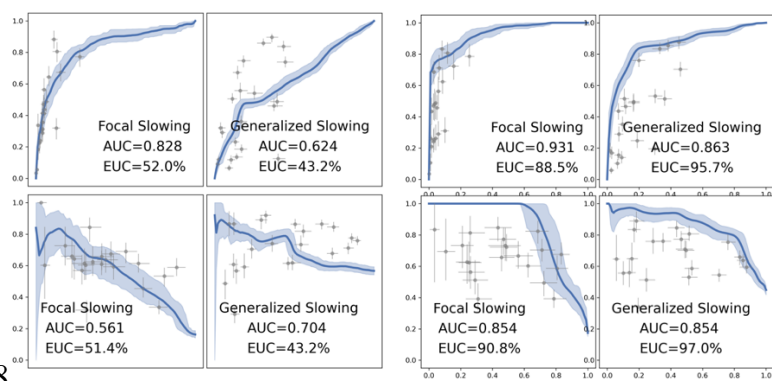




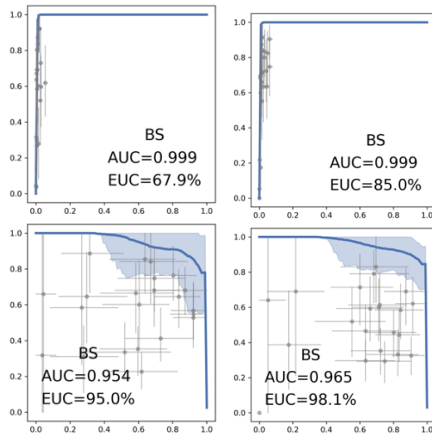
(A) Internal IIIC-Test dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue), Kaggle winner (yellow), SPaRCNet (green), and expert points with 95% CIs. Female results are shown at the top, and male results at the bottom; AUC-ROC Distribution by Model and Gender, where box plots showing the distribution of results from 10,000 bootstrap iterations, comparing the distribution of results across different age subgroups for each model



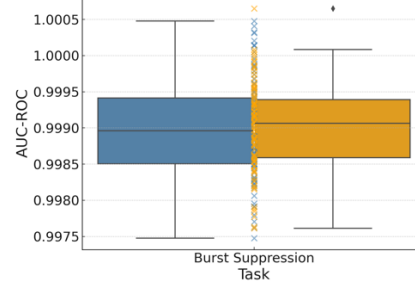
(B) Internal SN2-Test dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue), SpikeNet 2 (yellow) and expert points with 95% CIs. Female results are on the left and male results are on the right; AUC-ROC Distribution by Model and Gender, where the box plots show the distribution of results from 10,000 bootstrap iterations, comparing the distribution of results across different age subgroups for each model



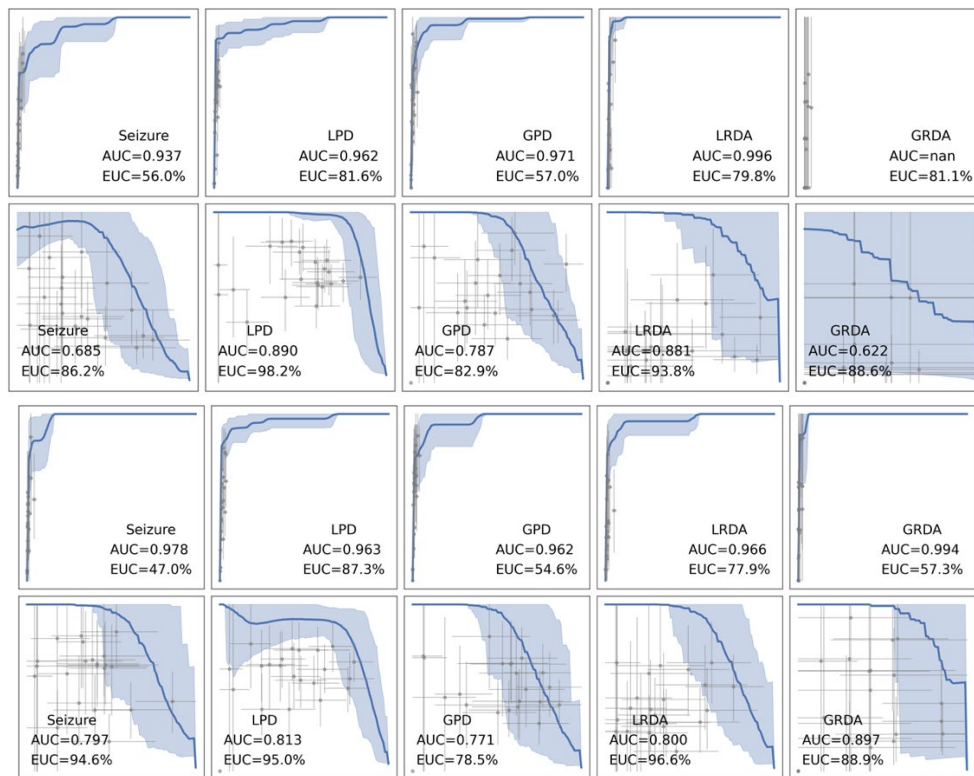
Slowing Localization



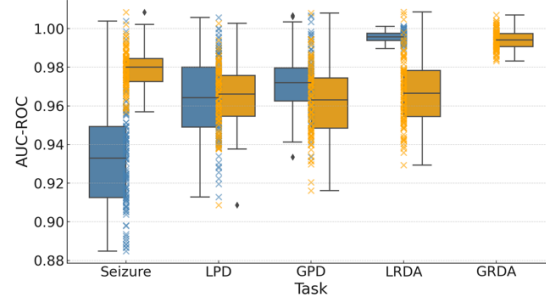
Morgoth Burst Suppression Results - AUC-ROC Distribution by Model and Gender



Burst Suppression Detection

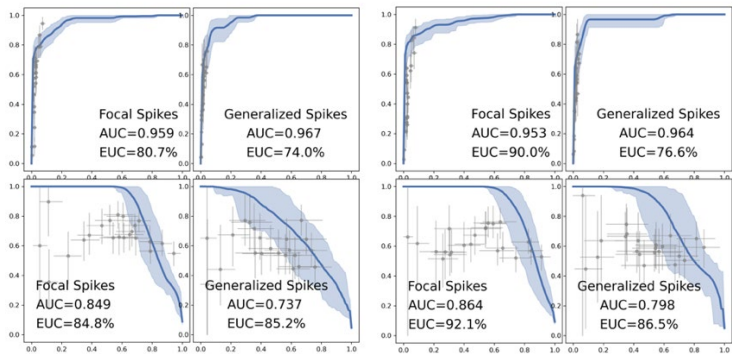


Morgoth IIIC Results - AUC-ROC Distribution by Model and Gender

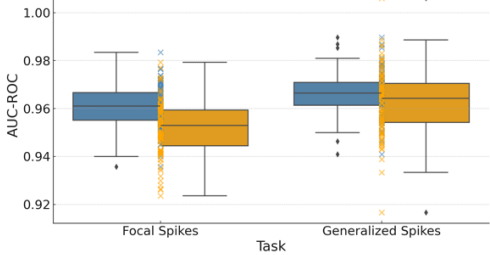


IIIC Classification

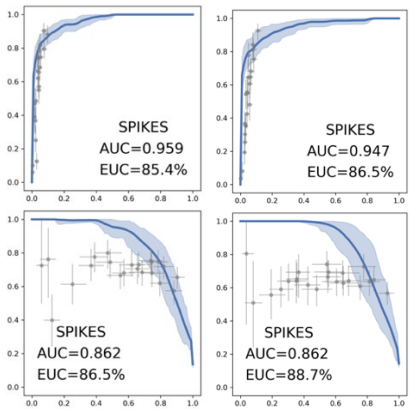
849



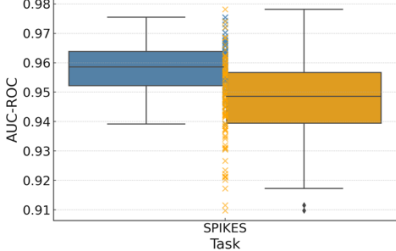
Morgoth Spike Detection Results - AUC-ROC Distribution by Model and Gender



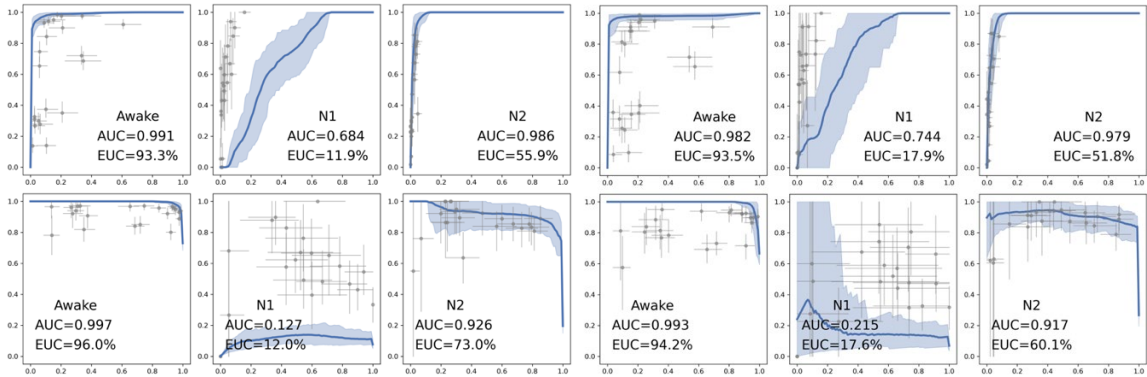
Spike Localization



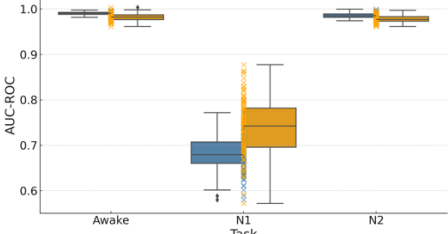
Morgoth Spike Results - AUC-ROC Distribution by Model and Gender



Spike Detection

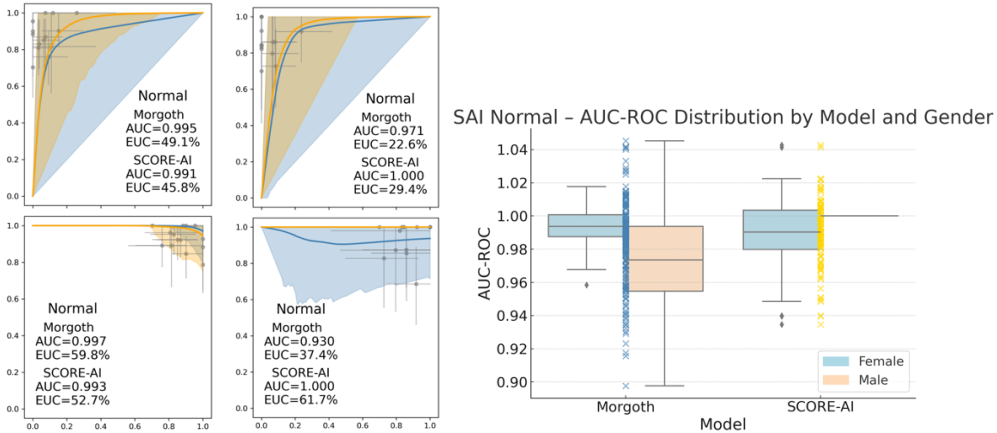


Morgoth Sleep Staging Results - AUC-ROC Distribution by Model and Gender

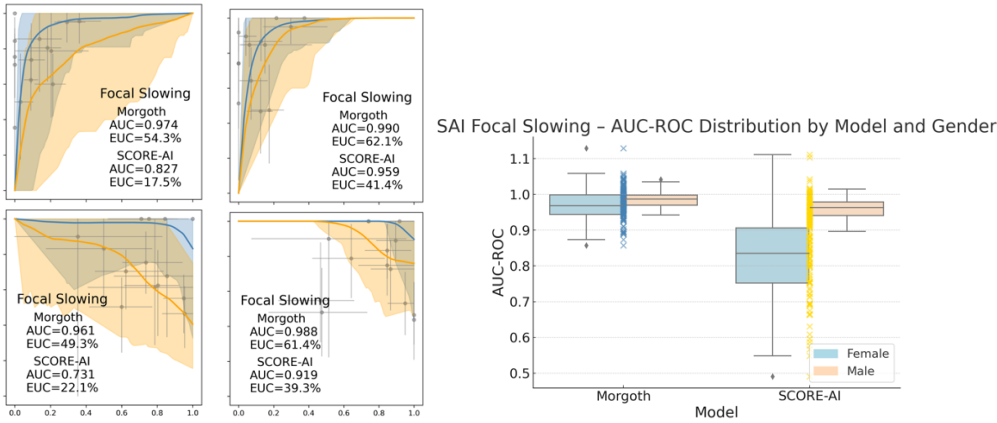


Sleep Staging

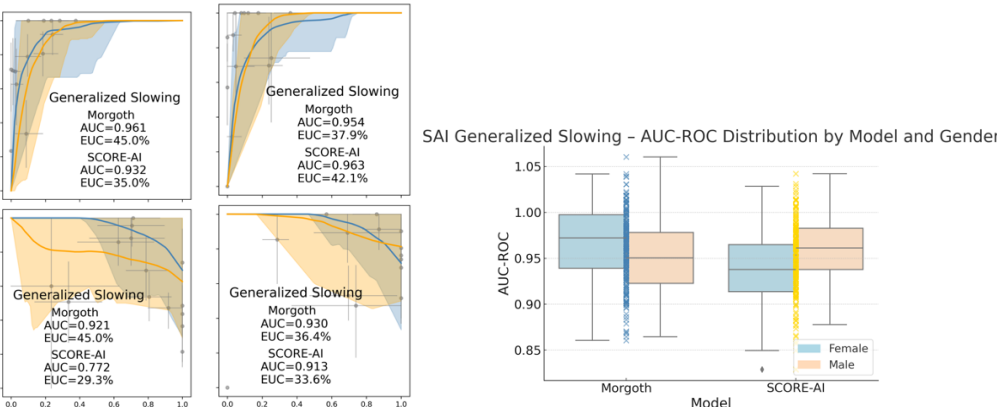
(C) Internal MoE dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue) and expert points with 95% CIs. Female results are on the left/top and male results are on the right/bottom.



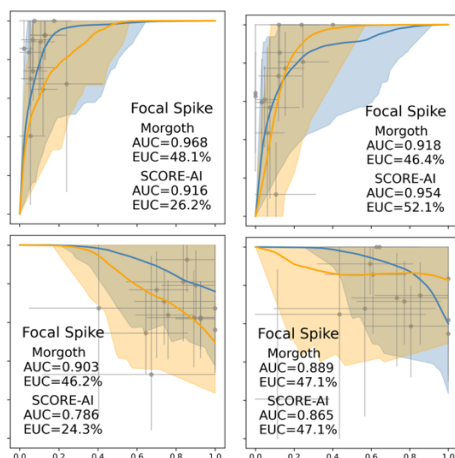
Normal vs Abnormal



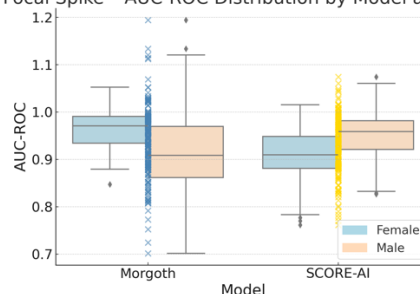
Focal Slowing Detection



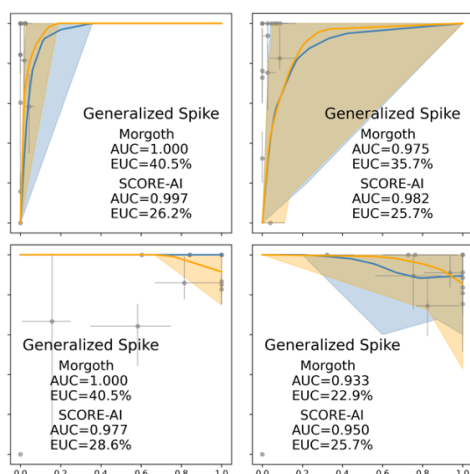
Generalized Slowing Detection



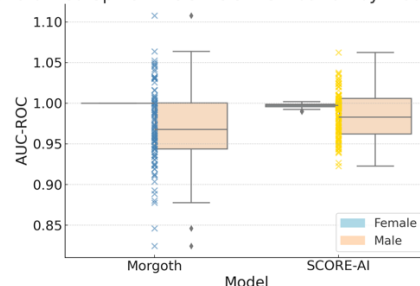
SAI Focal Spike - AUC-ROC Distribution by Model and Gender



Focal Spike Detection

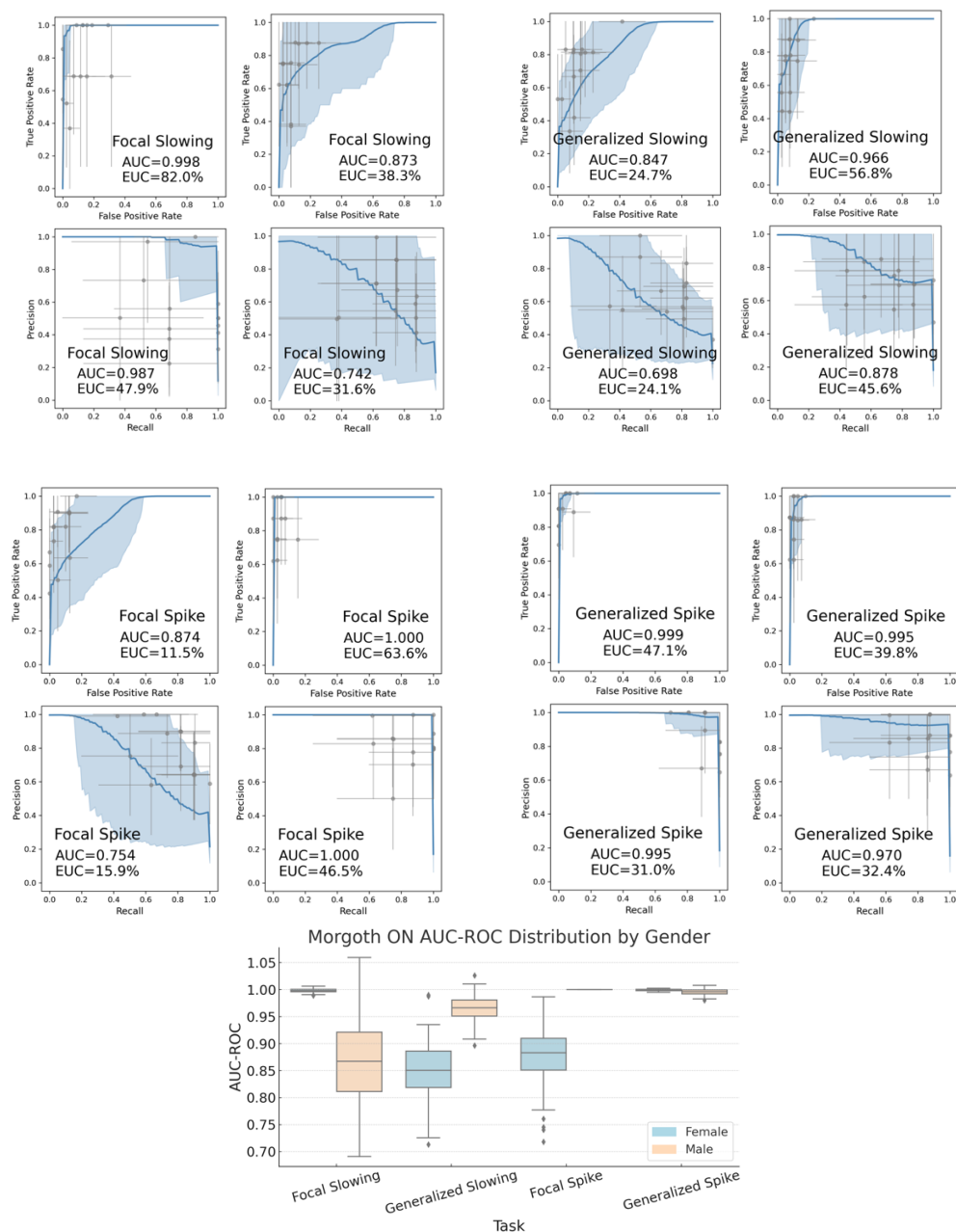


SAI Generalized Spike - AUC-ROC Distribution by Model and Gender



Generalized Spike Detection

(D) External SAI dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue), SCORE-AI (yellow) and expert points with 95% CIs. Female results are on the left and male results are on the right.



(E) External ON dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue) and expert points with 95% CIs. Female results are on the left and male results are on the right; AUC-ROC Distribution by Model and Gender, where box plots showing the distribution of results from 10,000 bootstrap iterations, comparing the distribution of results across different age subgroups for each model

Figure S22. Gender-specific Performance Results on Multi-expert Scored Datasets

899 11.2 Age group

900 As shown in Table S6, the six multi-expert annotated datasets were divided into 11 age groups.
901 To ensure that each group contains both positive and negative examples during testing, different
902 datasets were merged into different age groupings accordingly.

903 Overall results are shown in Figure S23 with age groups ordered in increasing manner from left
904 to right and top to bottom. All detailed results are shown in Table S7.

905 Analyzing whether model's AUROC depends on age were from two perspectives.

906 • Using regression (Figure S23)

907 Linear Regression - Start with a simple linear model: $\text{AUROC} \sim \text{Age}$. This will tell you if there's
908 a linear relationship and provide an easily interpretable coefficient showing how AUROC
909 changes per unit increase in age. Polynomial Regression - If you suspect non-linear relationships,
910 try quadratic or cubic terms: $\text{AUROC} \sim \text{Age} + \text{Age}^2 (+ \text{Age}^3)$. This captures potential U-shaped
911 or more complex age-performance relationships.

912 • Analyzing whether model's AUROC depends on age using distribution overlap (Table
913 S8)

914 We employed DeLong's test AUROC, AUPR, EUROC, EUCPR, values across different age
915 groups. DeLong's test is the recommended statistical method for AUC/EUC comparisons as it
916 properly accounts for the correlated nature of ROC curves derived from the same dataset and
917 provides exact p-values without requiring parametric assumptions about the underlying
918 distributions. We performed pairwise comparisons between all age groups and applied Bonferroni
919 correction for multiple comparisons. Statistical significance was defined as $p < 0.05$, and clinical
920 significance was considered as AUC/EUC differences ≥ 0.05 . This approach allows us to
921 rigorously assess whether observed performance differences across age groups represent true age-
922 dependent effects rather than random variation.

923 MORGOTH showed smaller differences between age groups compared to the baseline models,
924 demonstrating good generalization and robustness.

925 Based on the regression methods, across 27 feature detection tasks on five datasets, model
926 performance was found to be associated with age grouping in 5 tasks (Figure S23), including GPD
927 detection (Cubic, $p = 0.031$) and LRDA detection (Linear, $p = 0.023$) in the event-level IIIC-Test
928 dataset; Focal Slowing detection (Cubic = -0.031), LRDA detection (Linear, $p = 0.006$), and
929 Generalized Slowing detection (Linear, $p = 0.029$) in the EEG-level ON dataset.

930 For IIIC-Test dataset, age sensitivity analysis of three models in IIIC pattern detection reveals
931 significant model- and pattern-specific differences. Morgoth demonstrates the best overall age
932 robustness (20.9% average age sensitivity), particularly showing only 3.3% age sensitivity in
933 GRDA detection and significantly outperforming other models in LRDA detection (12.2% vs 46-
934 52%). In contrast, SPaRCNet exhibits the highest age sensitivity (30.9%), reaching 52.2% age

dependency in LRDA detection. From the IIIC pattern perspective, LRDA shows the highest age dependency (37.0%), followed by Seizure (36.2%) and LPD (31.5%), while GRDA demonstrates the best age robustness (8.7%). Notably, different models show significant variations in age sensitivity for the same IIIC pattern, with LRDA detection showing a 4-fold difference between Morgoth (12.2%) and SPaRCNet (52.2%).

These findings may have implications for clinical AI applications: GRDA detection can be directly applied across all age groups without age correction, and Morgoth should be the preferred choice for applications requiring cross-age-group consistency, while SPaRCNet and Kaggle Winner require implementation of age-specific performance calibration strategies in LRDA, Seizure, and LPD detection to ensure reliability and fairness of AI-assisted IIIC pattern detection across different age populations.

For the SN2-Test dataset, comparison of age-group performance between two spike detection models (Morgoth and SpikeNet2) reveals significant model-specific and age-dependent differences. Morgoth demonstrates age robustness across most age groups, achieving ROC AUC of 1.000 in all age groups except the >80 years group, with ROC EUC maintaining 100% consistency across the 1-50 years age range. In contrast, SpikeNet2 exhibits stronger age sensitivity, particularly in pediatric and adolescent groups (1-30 years), where its ROC EUC drops to 22.4%-59.4%, indicating reduced performance in younger populations. Both models maintain good age consistency in PR AUC metrics (≥ 0.990 in most age groups), but the PR EUC metric reveals more nuanced age-related performance patterns. The Morgoth model's PR EUC fluctuates considerably across age groups (47.8%-87.3%), peaking in middle-aged to older adult groups (40-60 years) while being relatively lower in young adult (20-40 years) and elderly groups (60-70 years). SpikeNet2 shows particularly poor PR EUC performance in younger groups, with only 22.4% in the 20-30 years group, suggesting that this model may have overfit to adult spike characteristics and lacks generalization capability for pediatric and adolescent spike detection. Notably, both models show identical performance in the 30-40 and 40-50 years groups, while Morgoth demonstrates better performance stability in populations over 50 years. These findings indicate that Morgoth possesses superior cross-age-group generalization, whereas SpikeNet2 exhibits age bias, particularly insufficient spike detection performance in pediatric and adolescent populations

For the MoE dataset, among the 14 EEG pattern automatic detection models, the Generalized Slowing detection exhibited the strongest age sensitivity (overall significant proportion 51.6%), with metric 2 achieving 73.3% significant proportion and an extremely low average p-value of 0.063. Sleep-related pattern detection models (Sleep-N1 and Sleep-N2) also demonstrated significant age dependency, with overall significant proportions of 33.4% and 26.6% respectively, indicating that these models' detection performance exhibits age bias and may produce systematic detection differences across different age groups. In contrast, the Spike detection model demonstrated ideal age robustness (0.0% significant proportion), while GPD and LPD intensive care pattern detection models were largely unaffected by age (1.7%-5.0%), suggesting that these models successfully learned age-independent pathological features and possess good cross-age-group generalization capability. From the perspective of metric types, mean-type metrics (ROC AUC, PR AUC) showed higher age sensitivity in model predictions (20.5%-21.9%) compared to

percentage-type metrics (ROC EUC and PR EUC, 11.4%-13.4%), indicating that different types of output metrics exhibit varying degrees of age generalization capability.

For the SAI dataset, age sensitivity comparison analysis between Morgoth and SCORE-AI on EEG-level tasks reveals significant task-specific differences. Morgoth demonstrates slightly superior overall age robustness (6.5% vs 8.0%), outperforming SCORE-AI in Generalized Slowing detection (12.5% vs 22.5%), while both models exhibit perfect age robustness (0.0%) in Generalized Spike detection. From the EEG task perspective, Generalized Slowing shows the highest age dependency (17.5%), becoming the primary source of age sensitivity for both models, while Normal state detection, Focal Slowing, and Generalized Spike detection all demonstrate good age robustness (0-3.8%). Notably, ROC-AUC metrics show 0.0% age sensitivity across all task-model combinations, while EUC metrics display more age-related variation, particularly with SCORE-AI's ROC-EUC reaching 60.0% age sensitivity in Generalized Slowing detection. These findings indicate that on EEG-level detection tasks Morgoth demonstrates superior age robustness compared to complex IIC pattern detection (average 7.3% vs 25.8%), providing important guidance for clinical AI applications: Generalized Spike and Normal state detection can be directly applied across all age groups, Morgoth should be preferred for applications requiring cross-age-group consistency, while SCORE-AI requires implementation of age-specific performance calibration strategies in Generalized Slowing detection to ensure reliability of AI-assisted EEG diagnosis across different age populations.

For the ON dataset, generalized spike detection exhibits the highest age sensitivity (32.5%), primarily driven by extremely high age dependency in EUC metrics (PR-EUC 60.0%, ROC-EUC 70.0%), while its AUC metrics demonstrate complete age robustness (0.0%). In contrast, generalized slowing detection shows perfect age robustness (0.0%), with all metrics showing no significant differences across age groups.

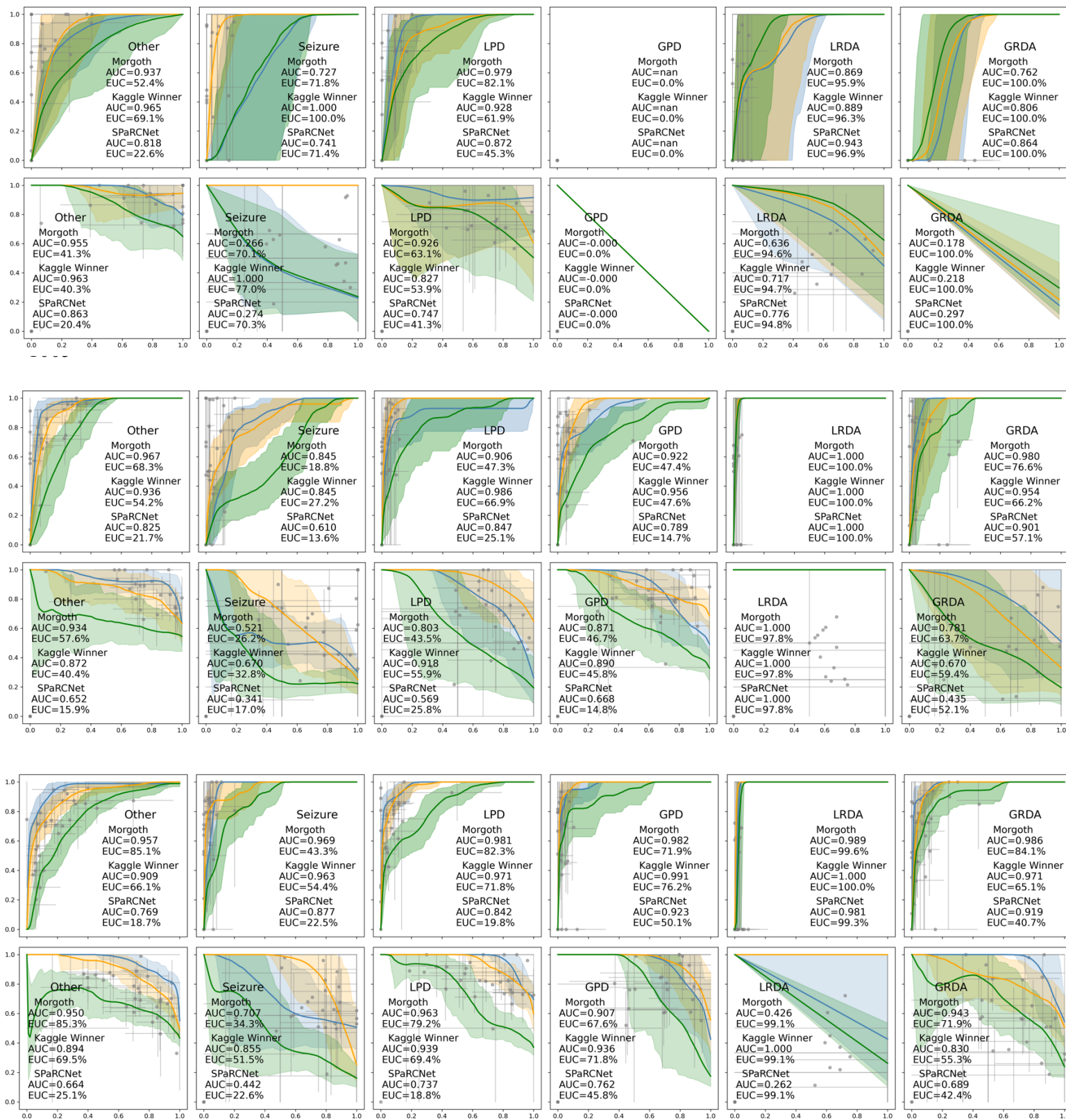
1001

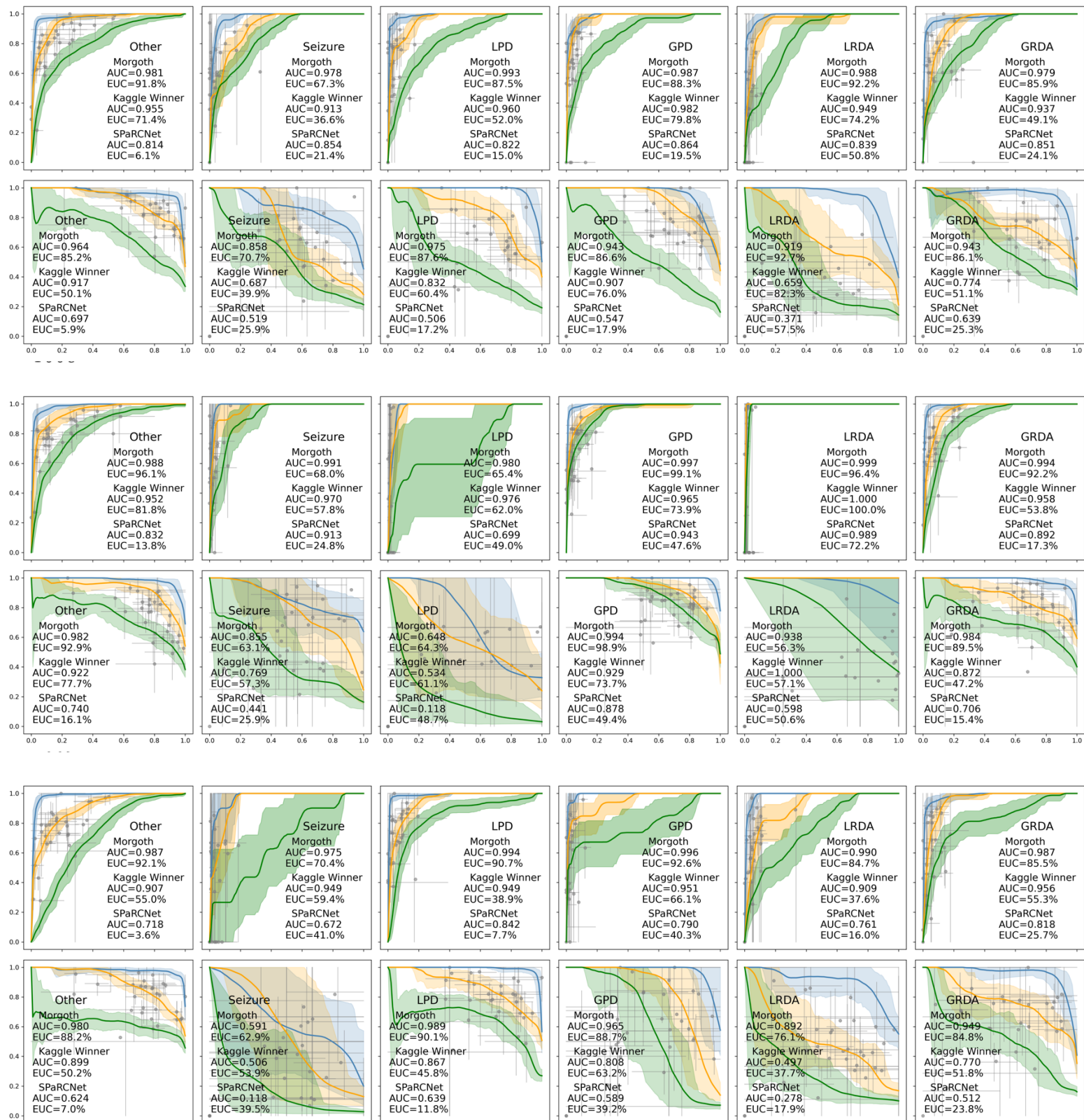
Table S6 Age Groups of Multi-expert Annotated Datasets

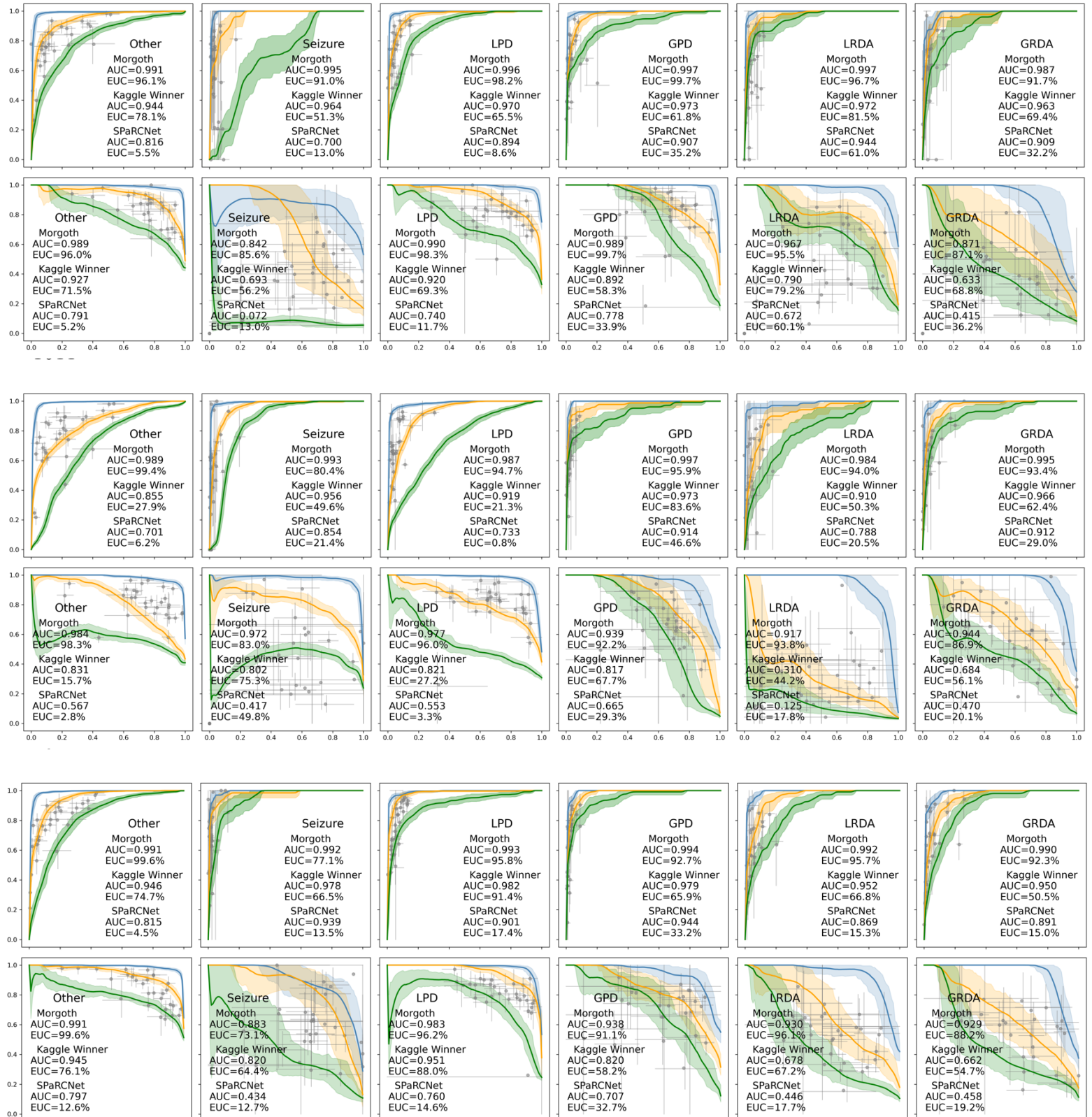
Age Group*	IIC	Sub Group	SN2	Sub Group	SAI	Sub Group	MoE	Sub Group	ON	Sub Group
0-1	22	0-1	7	0-1	1	0-10	28	0-20	3	0-10
1-10	51	1-10	34	1-10	13		92		25	
10-20	41	10-20	36	10-20	26	10-30	101		19	10-20
20-30	49	20-30	19	20-30	12		238	20-30	7	20-40
30-40	46	30-40	12	30-40	9	30-50	219	30-50	6	
40-50	46	40-50	9	40-50	8		281		6	40-70
50-60	117	50-60	19	50-60	7	50-70	427	50-60	13	
60-70	140	60-70	36	60-70	14		440	60-70	6	
70-80	143	70-80	23	70-80	6	>70	363	>70	11	>70
80-90	78	80-90	12	>80	3		188		1	
>90	14	>90	2		1		23		1	

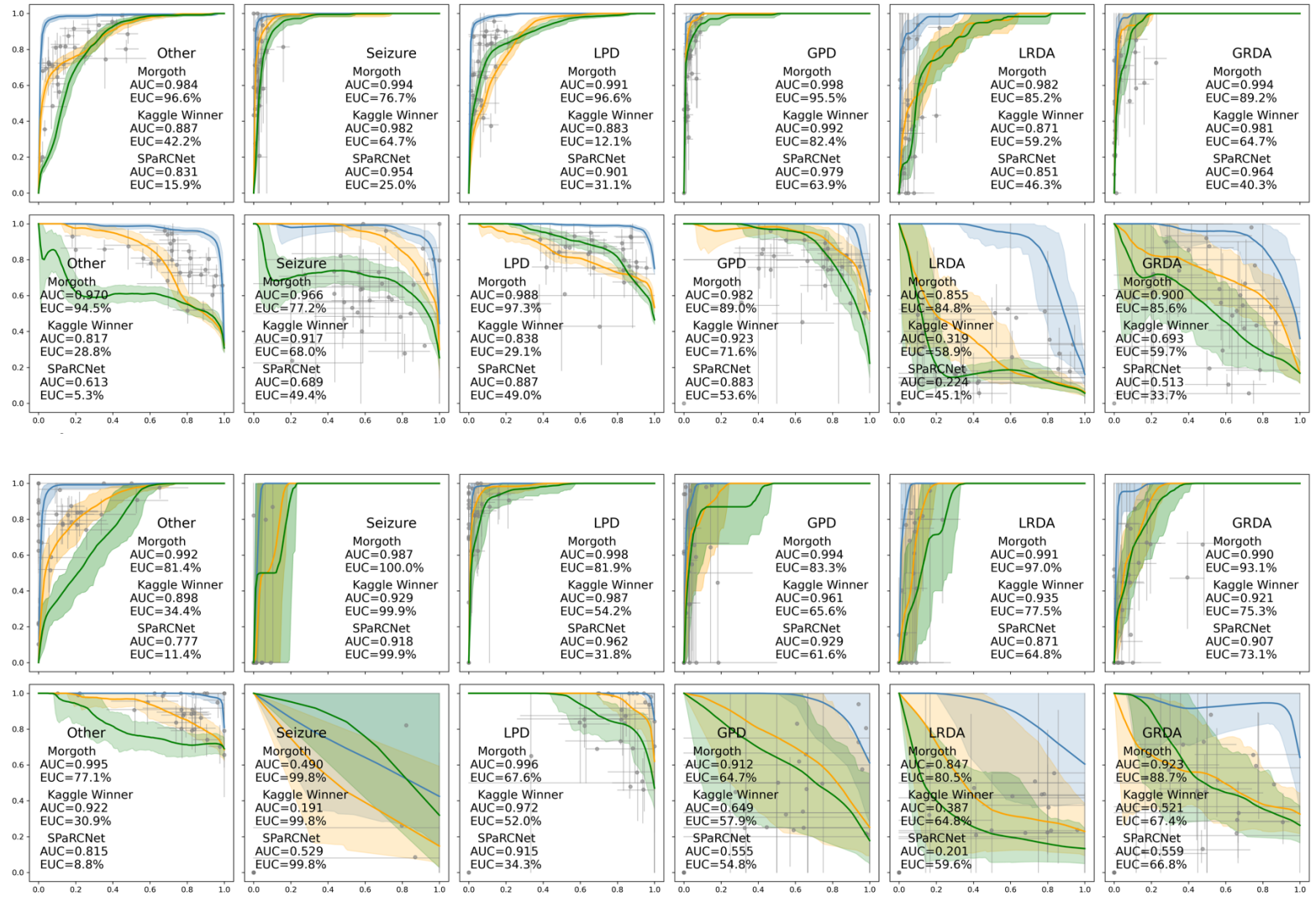
* All intervals are left-closed and right-open.

1004



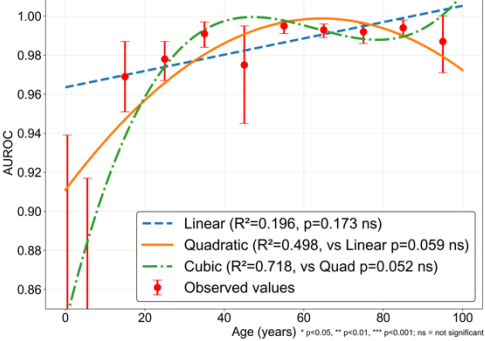




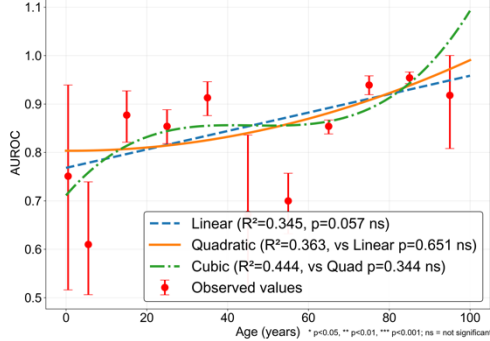


1016

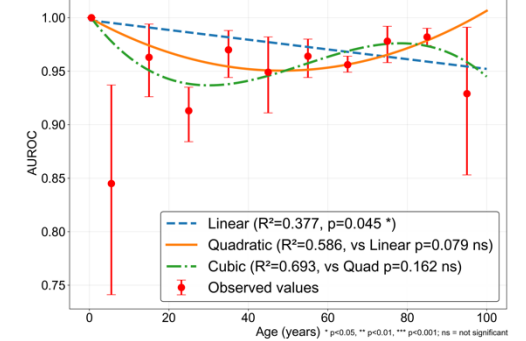
IIIC-Test Seizure - Morgoth Performance vs Age - Regression Analysis

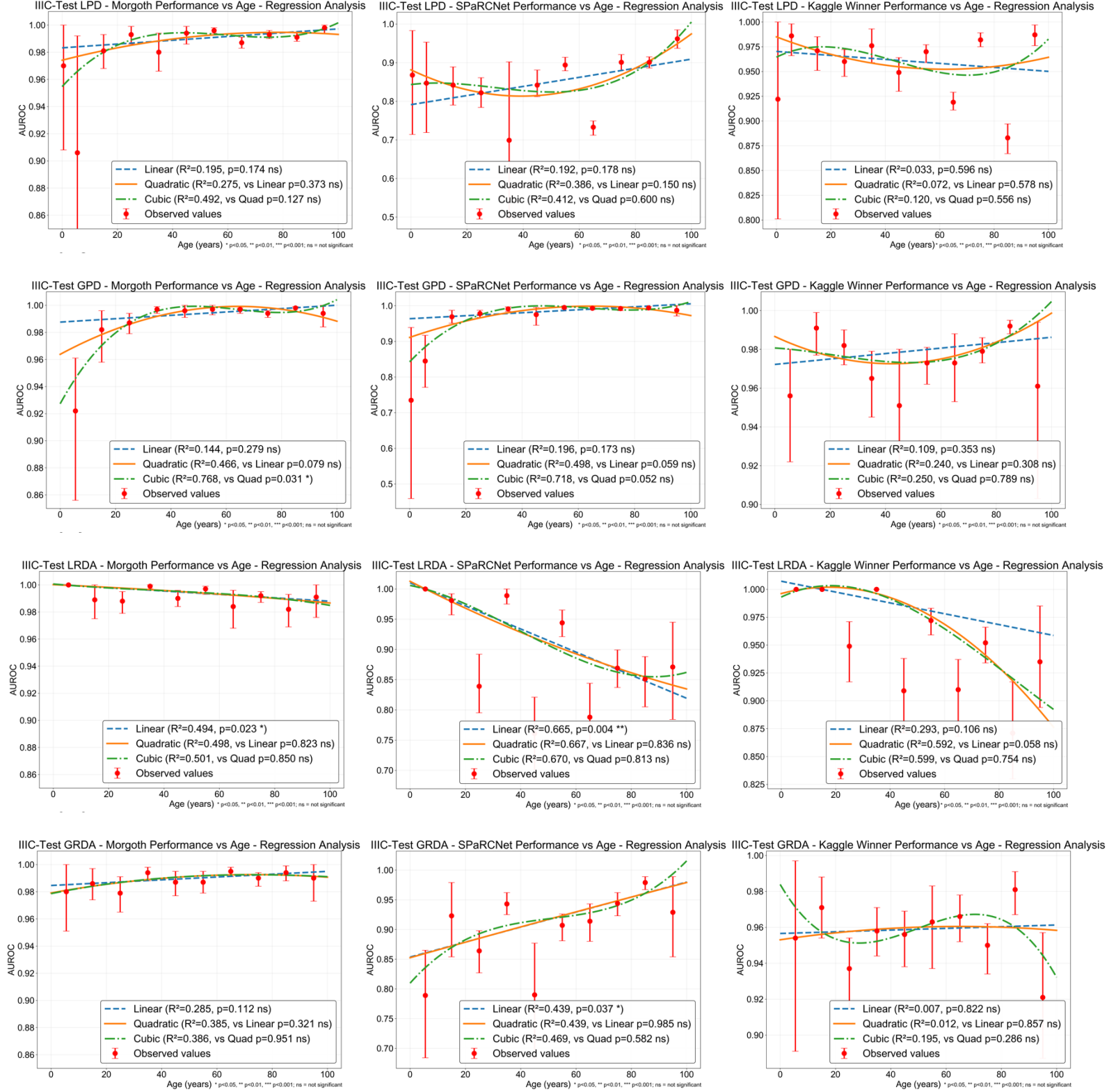


IIIC-Test Seizure - SPaRCNet Performance vs Age - Regression Analysis



IIIC-Test Seizure - Kaggle Winner Performance vs Age - Regression Analysis

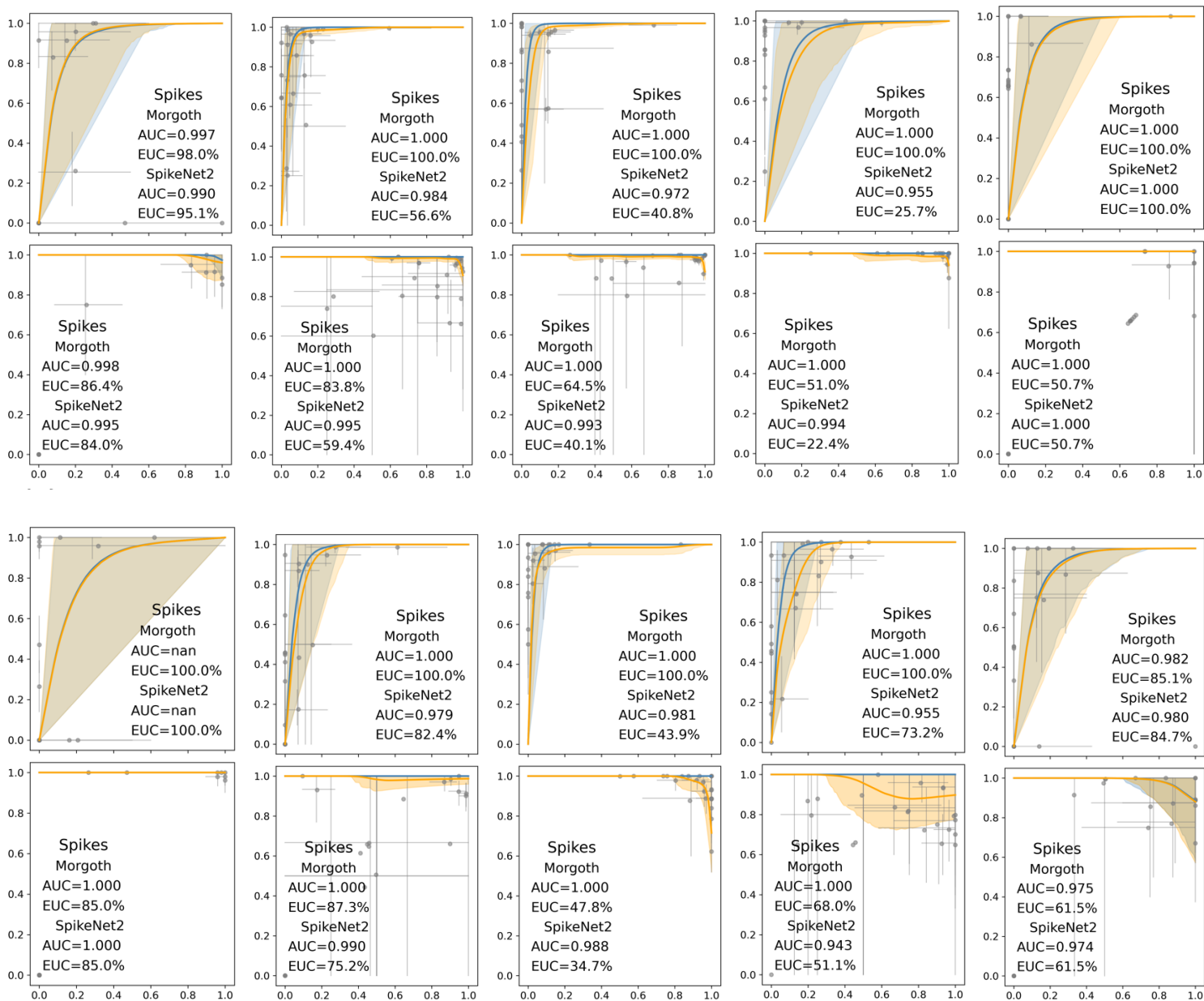


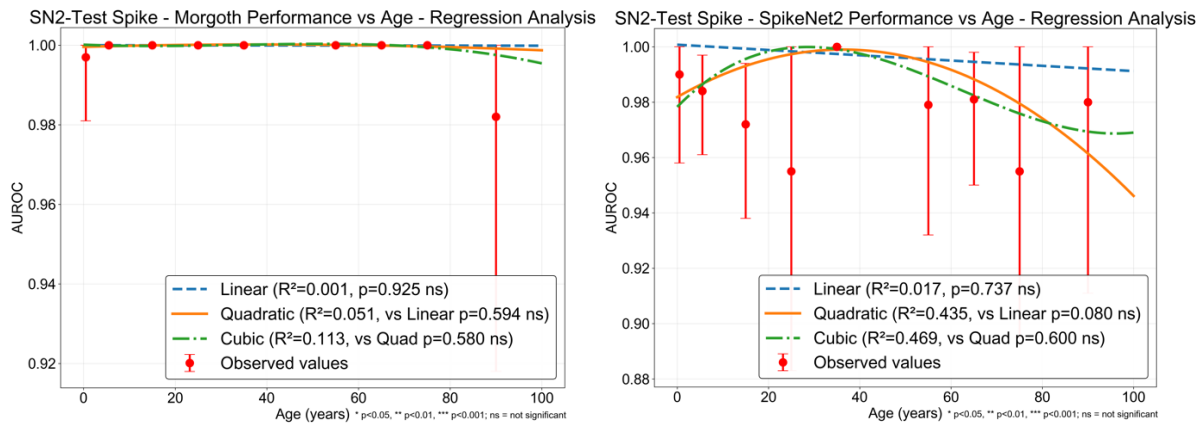


1022 (A) Internal IIIC-Test dataset: Average ROC curves (odd rows) and PR curves (even rows) for
 1023 Morgoth (blue), Kaggle winner (yellow), SPARCNet (green), and expert points with 95% CIs.
 1024 The age groups from top to bottom are as follows: [0, 1), [1, 10), [10, 20), [20, 30), [30, 40),

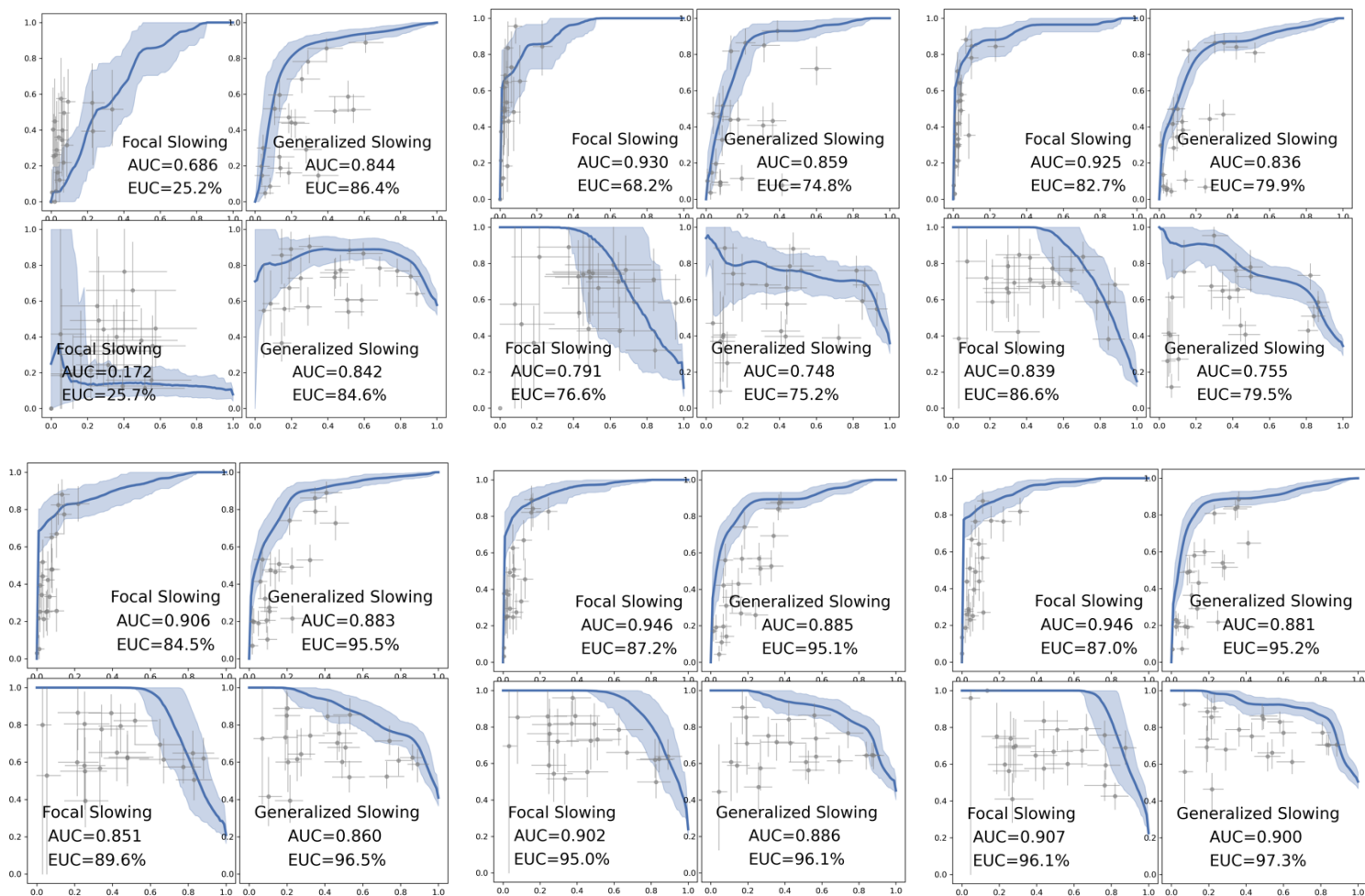
1025 [40, 50), [50, 60), [60, 70), [70, 80), [80, 90), and >90; Regression between age and AUC-
 1026 RPC - if age significantly affect model performance. * $p<0.05$, ** $p<0.01$, *** $p<0.001$; ns =
 1027 not significant.

1028



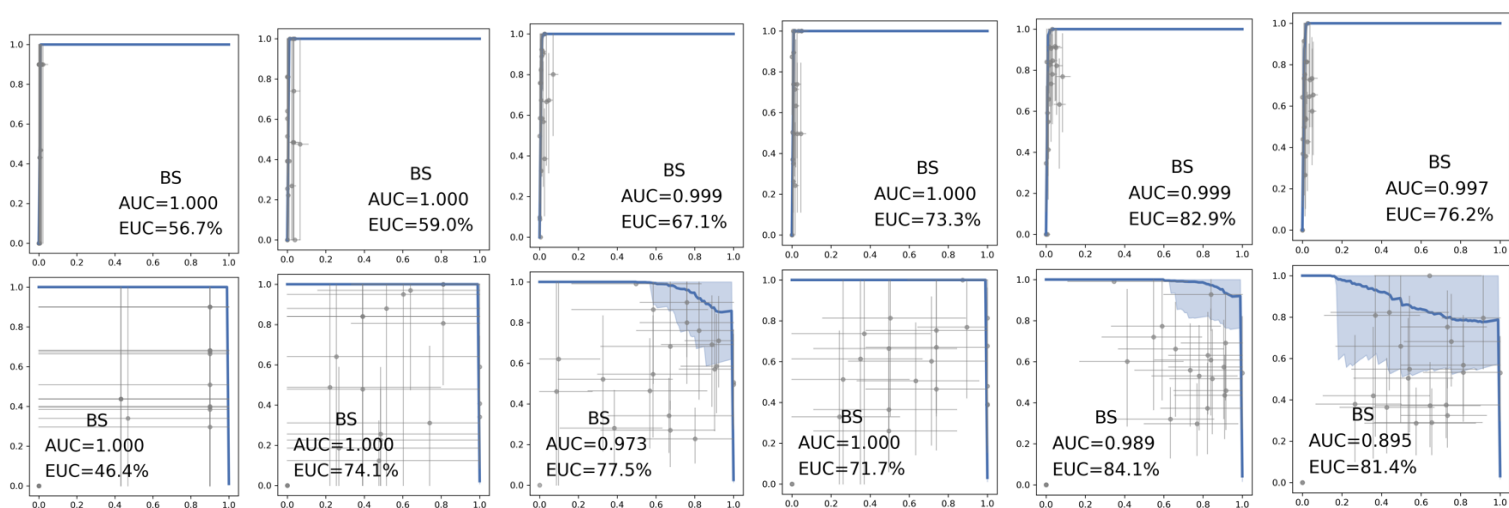


(B) Internal SN2-Test dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue), SpikeNet 2 (yellow) and expert points with 95% CIs. The age groups from left to right and top to bottom are: [0, 1), [1, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80), and >80. Regression between age and AUC-RPC - if age significantly affect model performance. * $p<0.05$, ** $p<0.01$, *** $p<0.001$; ns = not significant.



1040

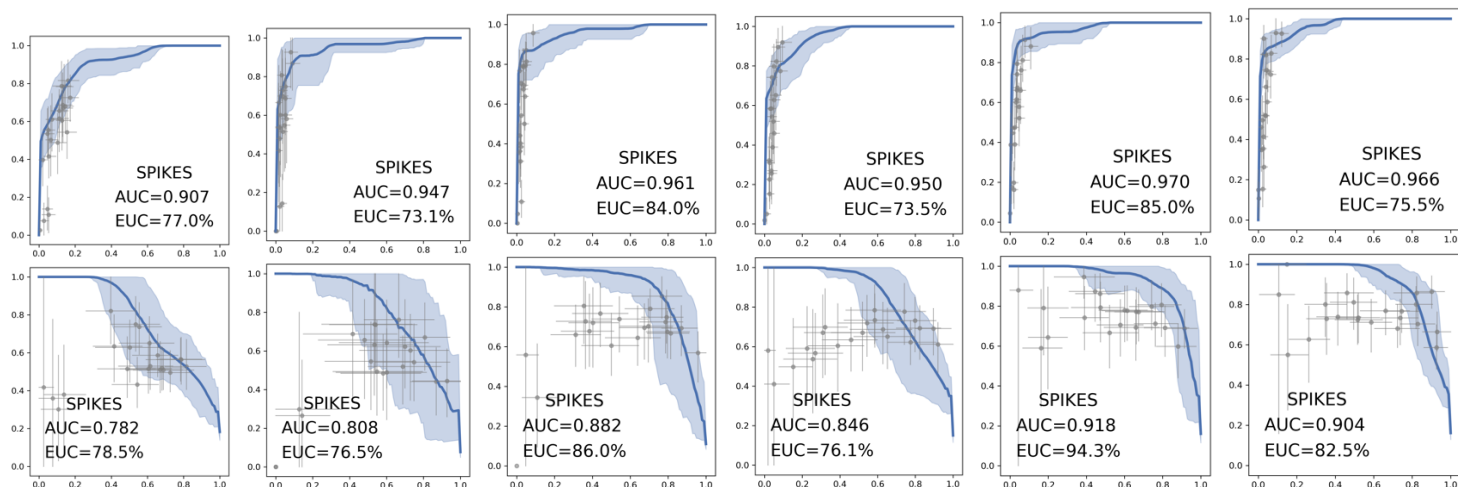
Slowing Localization



1042

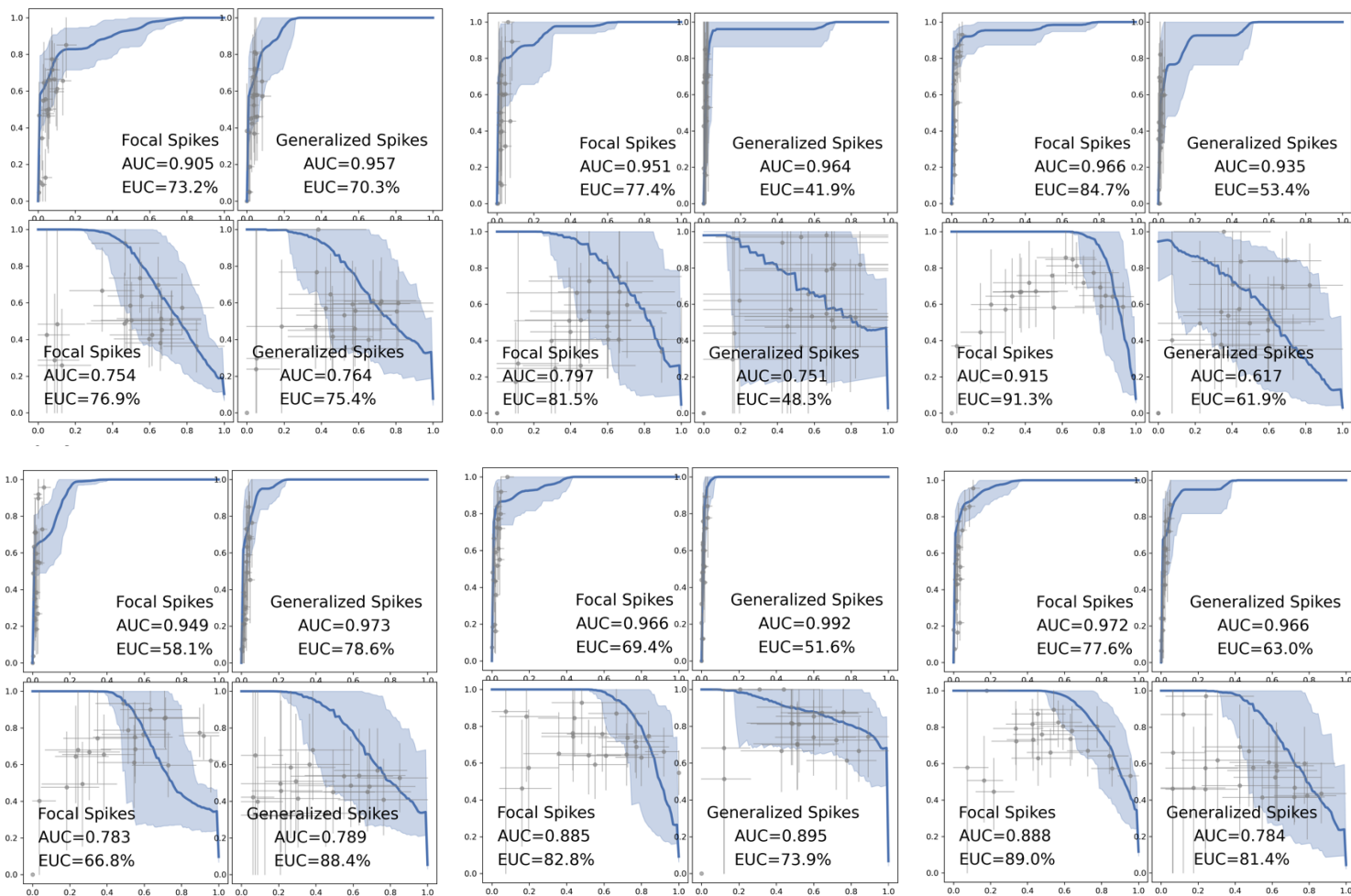
Burst Suppression Detection

1043

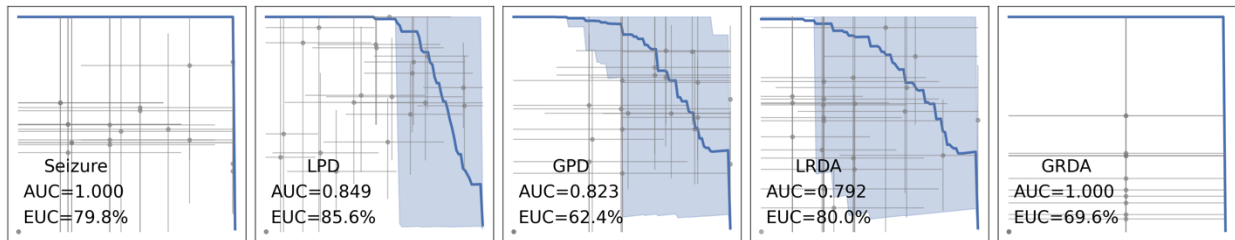
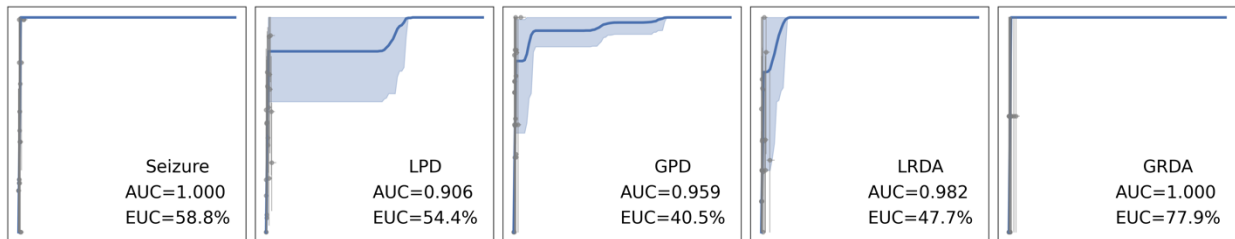
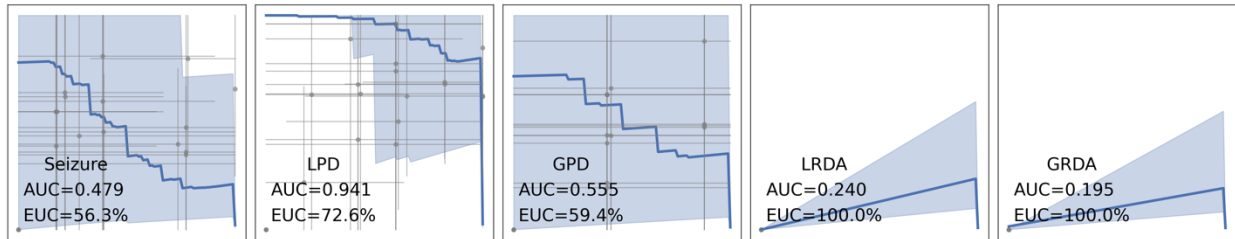
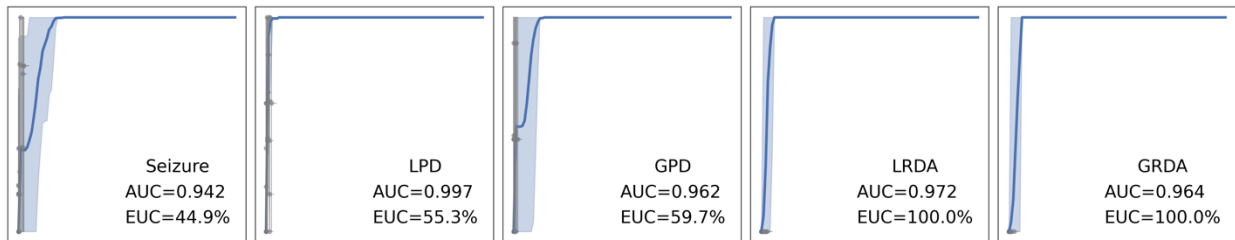
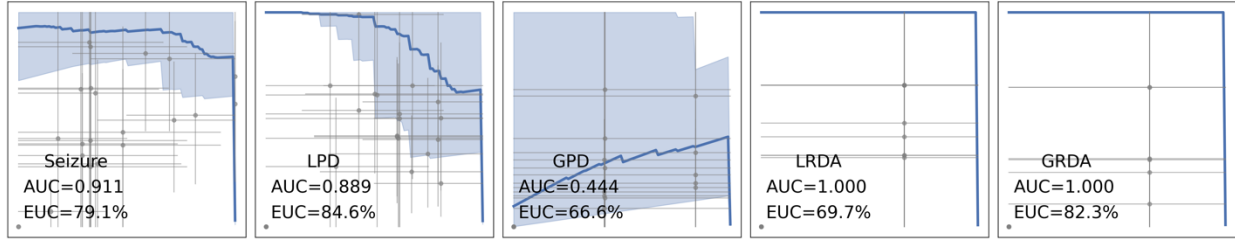
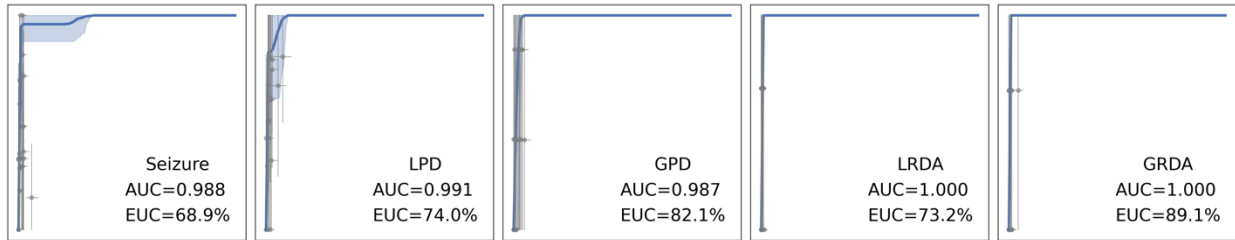


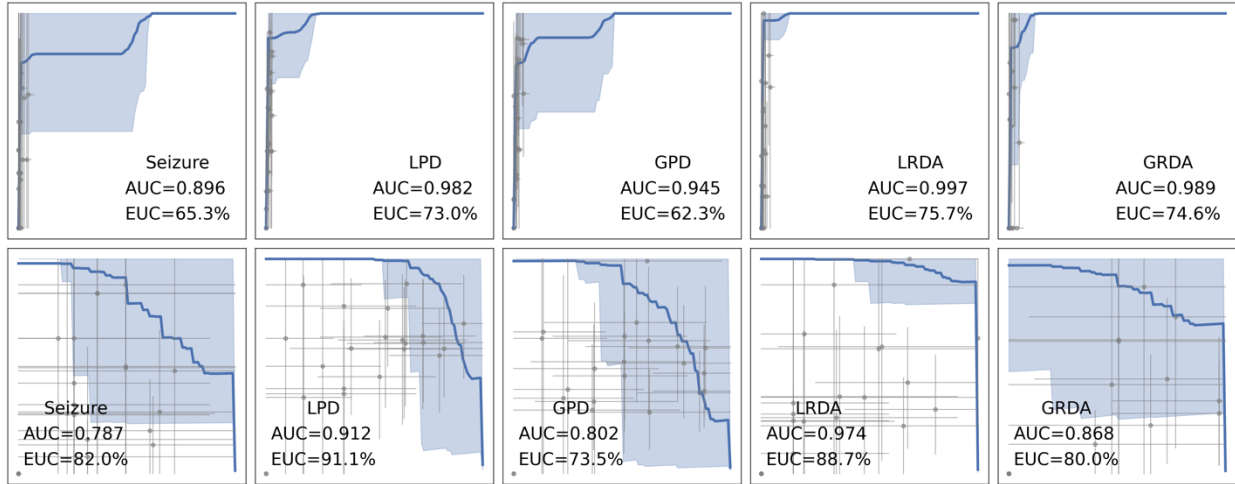
1045

Spike Detection

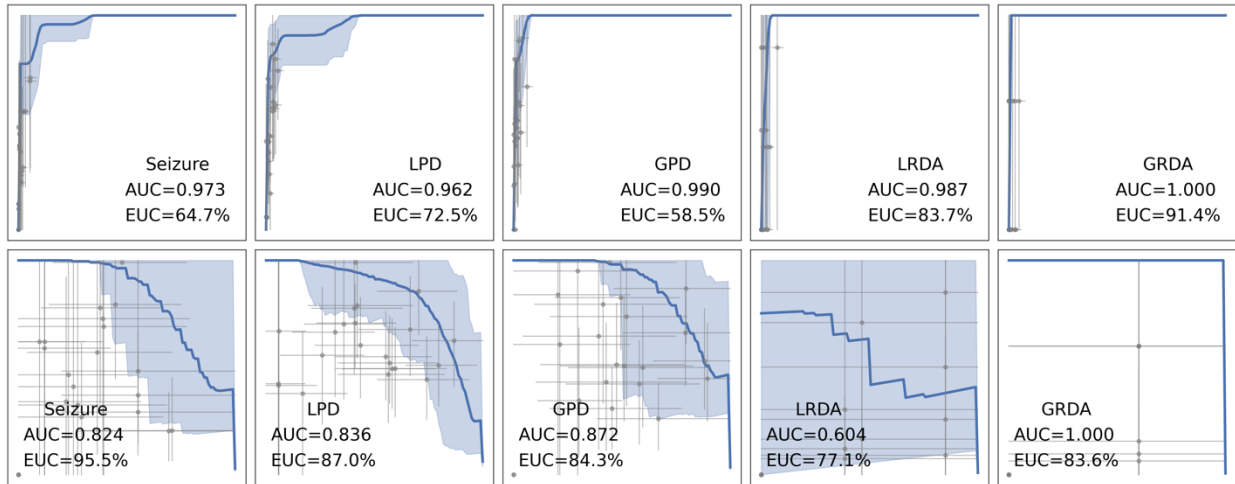


Spike Localization

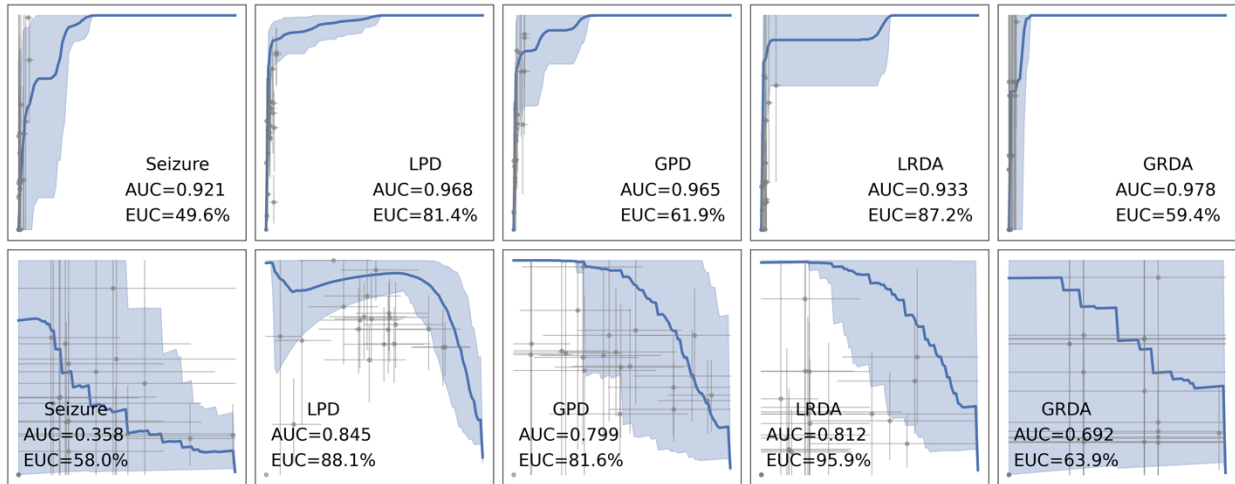




1053



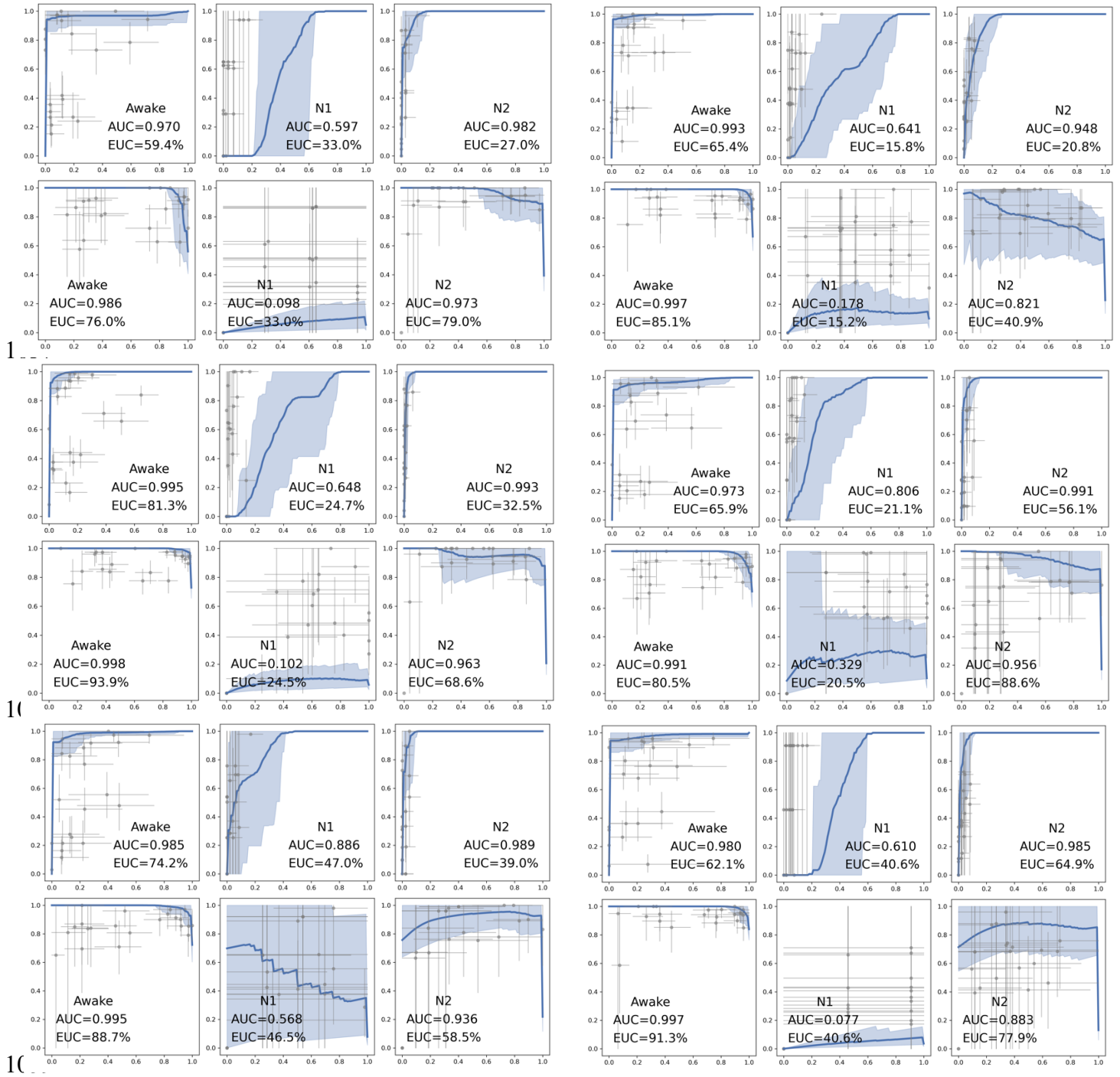
1054



1055

1056

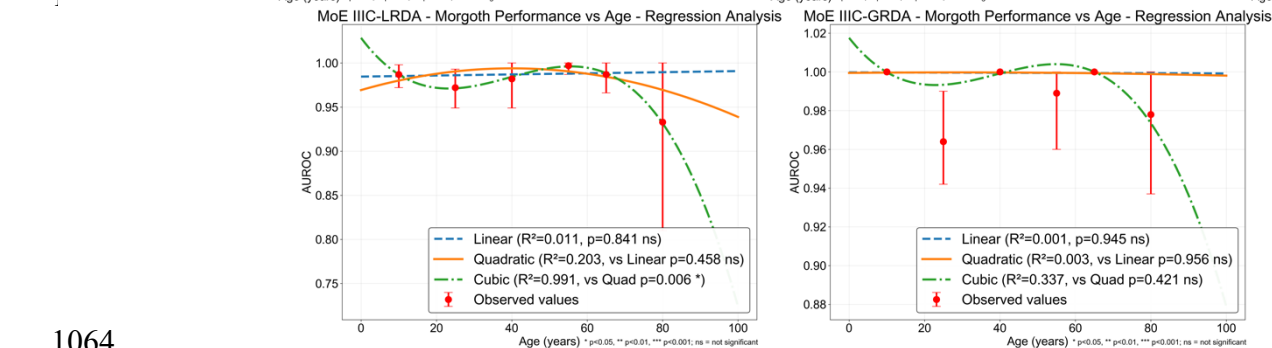
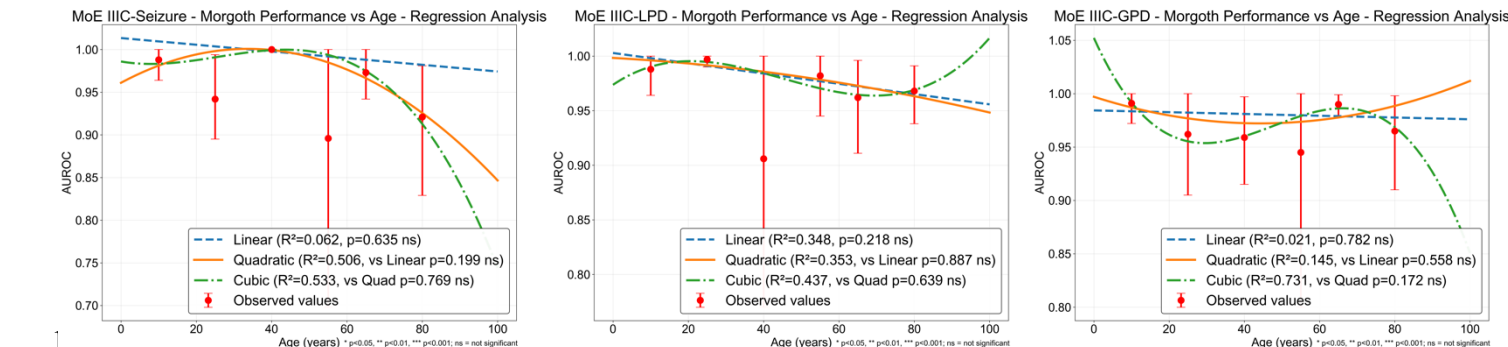
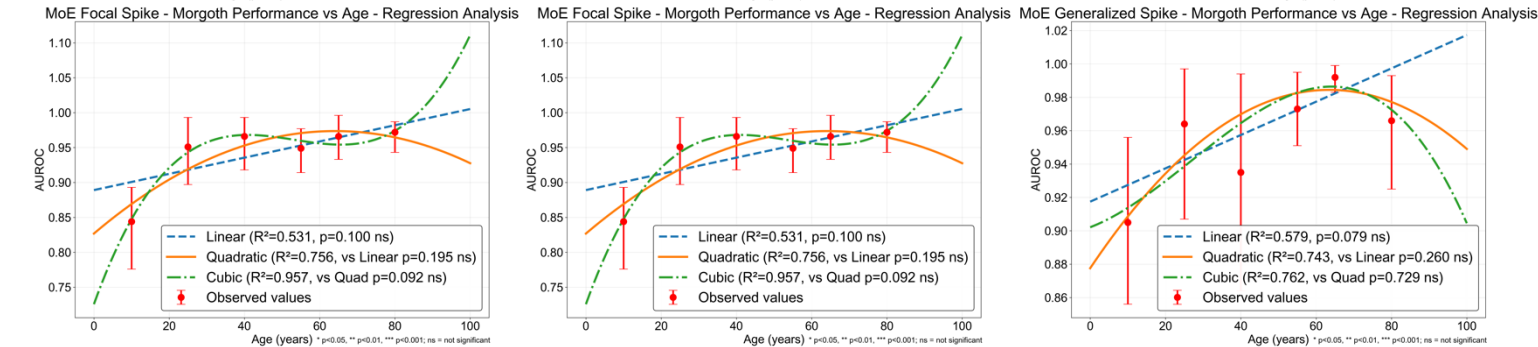
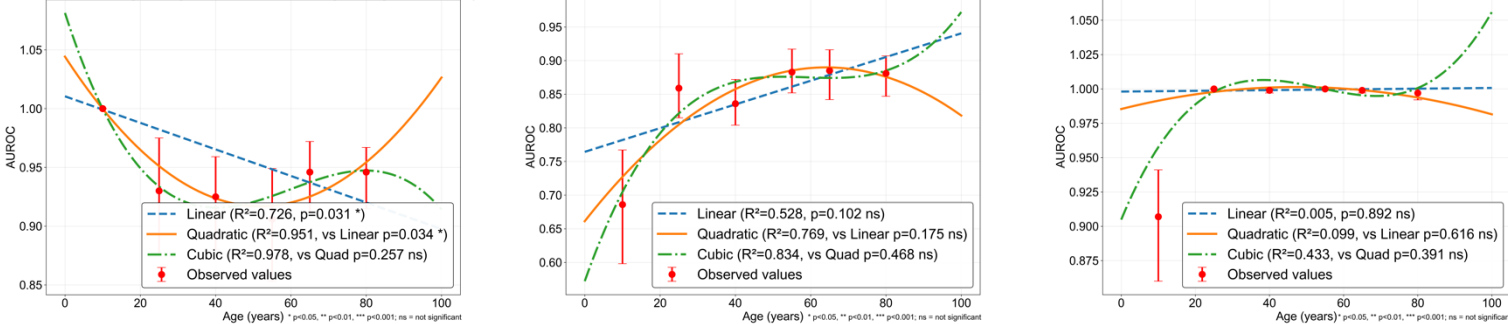
IIIC Classification

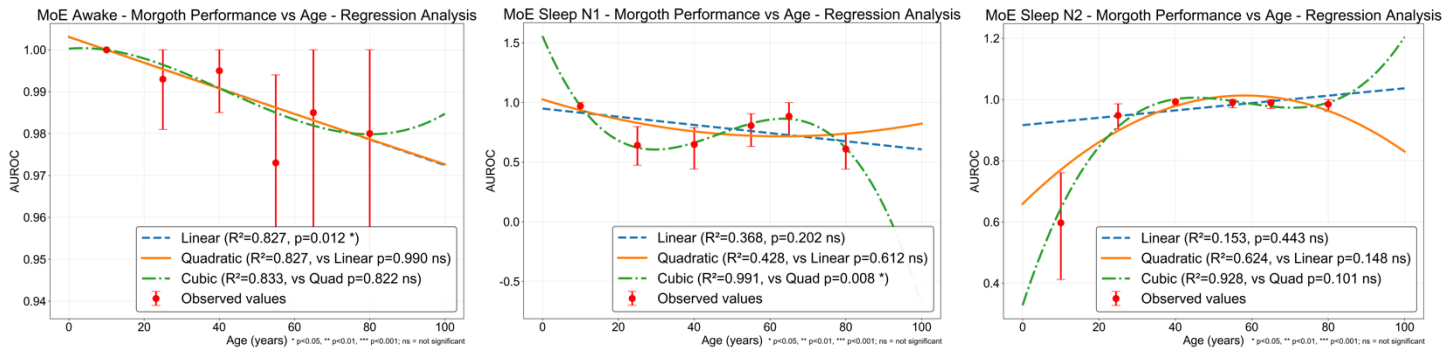


1060

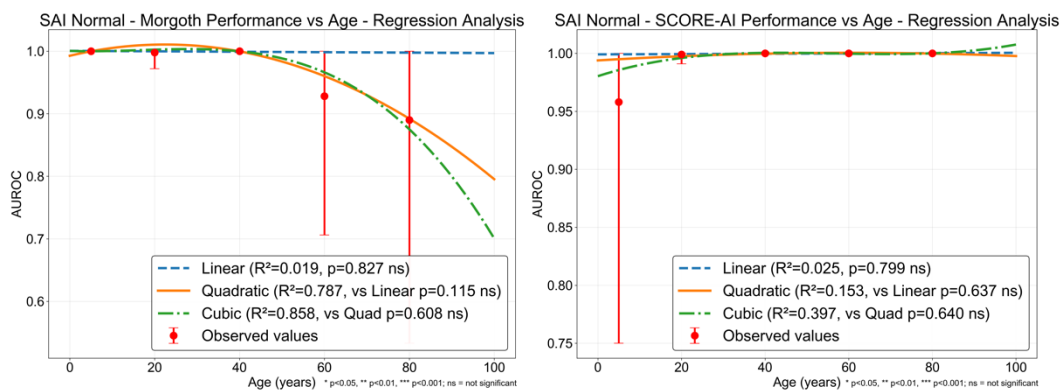
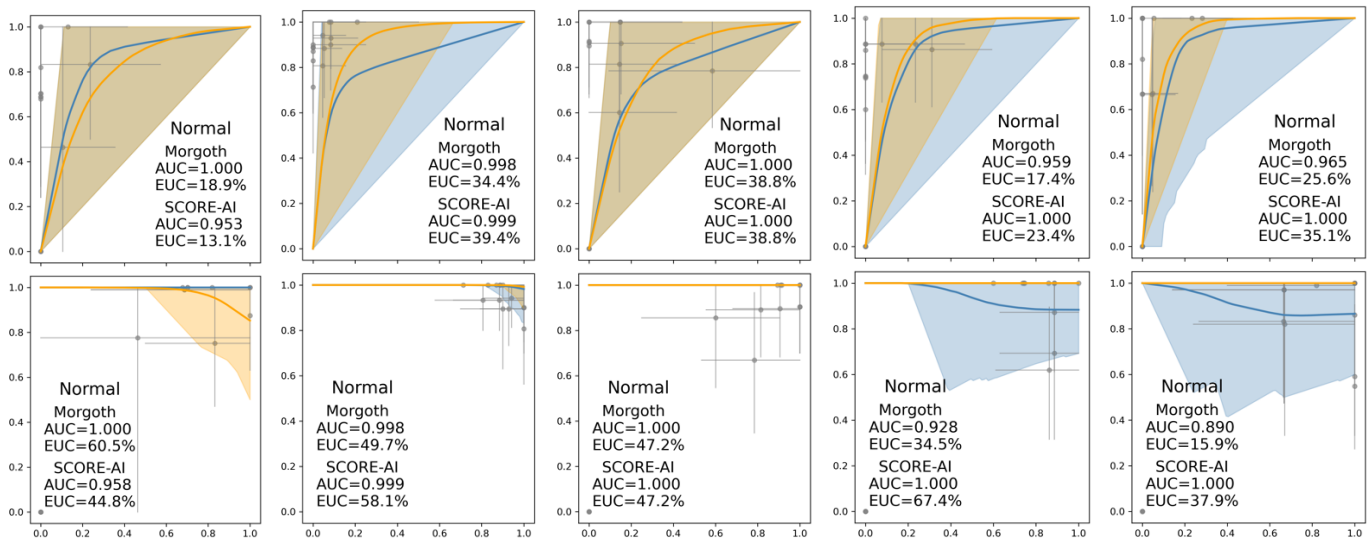
Sleep Staging

MoE Focal Slowing - Morgoth Performance vs Age - Regression Analysis MoE Generalized Slowing - Morgoth Performance vs Age - Regression Analysis MoE Burst Suppression - Morgoth Performance vs Age - Regression Analysis

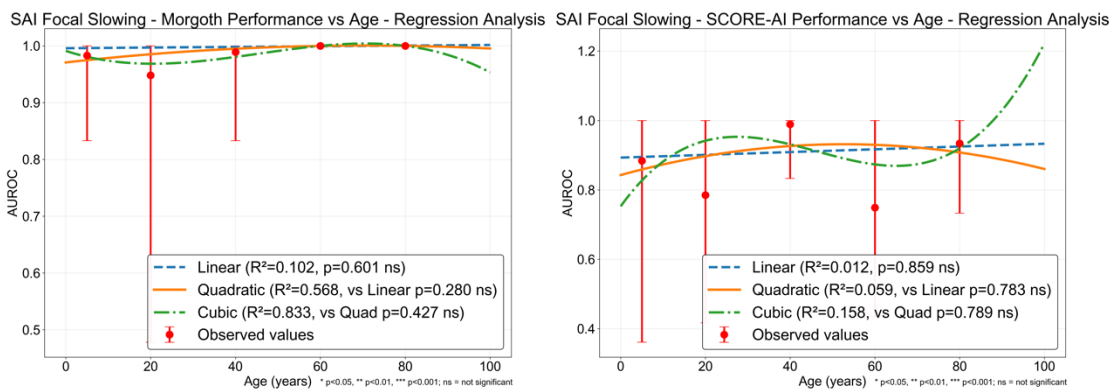
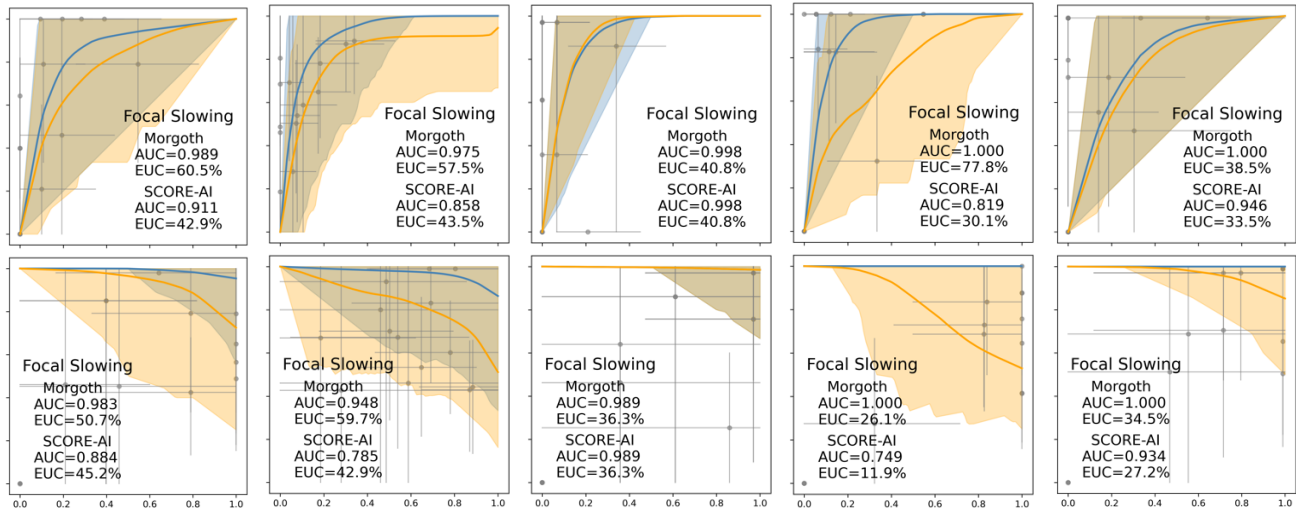




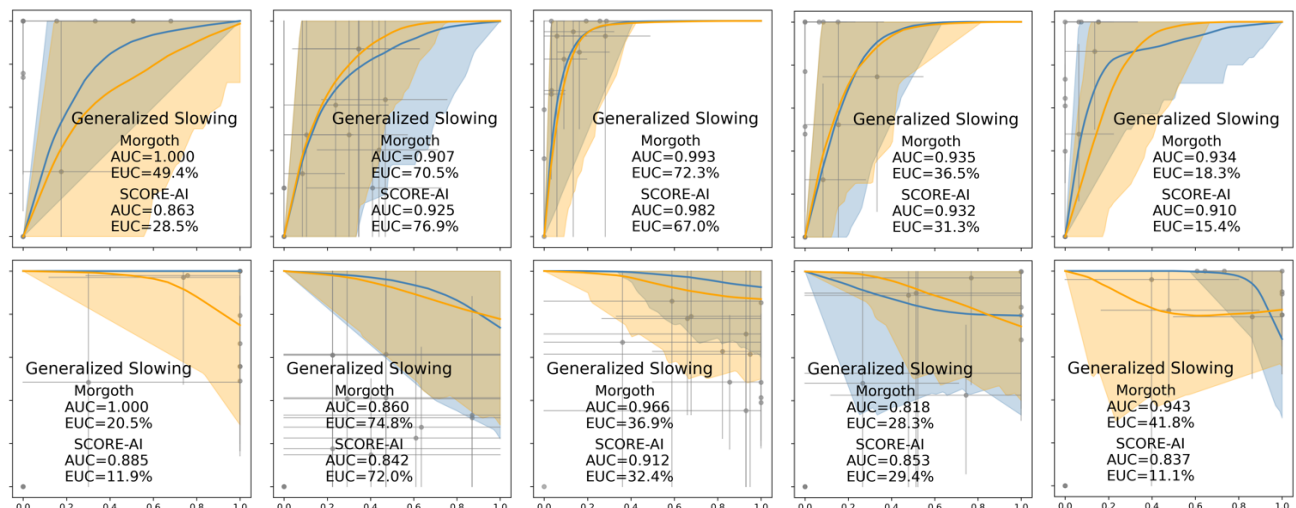
(C) Internal MoE dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue) and expert points with 95% CIs. The age groups from left to right and top to bottom are: [0, 20), [20, 30), [30, 50), [50, 60), [60, 70), and >70. Regression between age and AUC-RPC - if age significantly affect model performance. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ns = not significant.

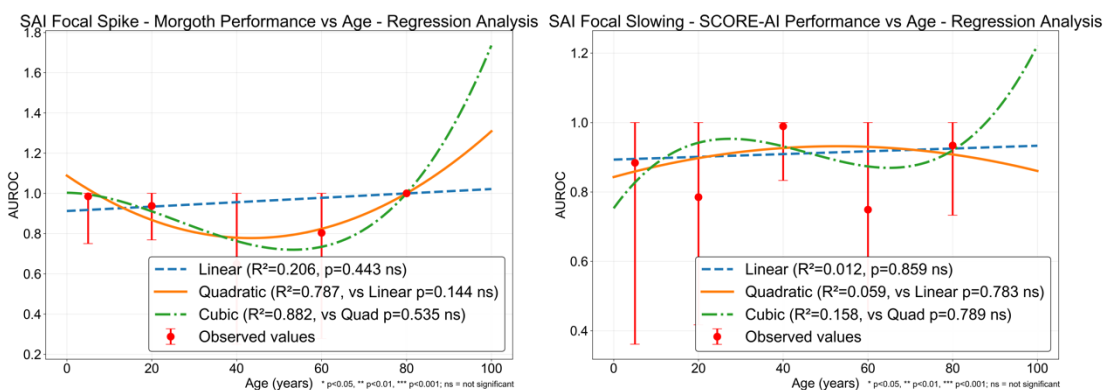
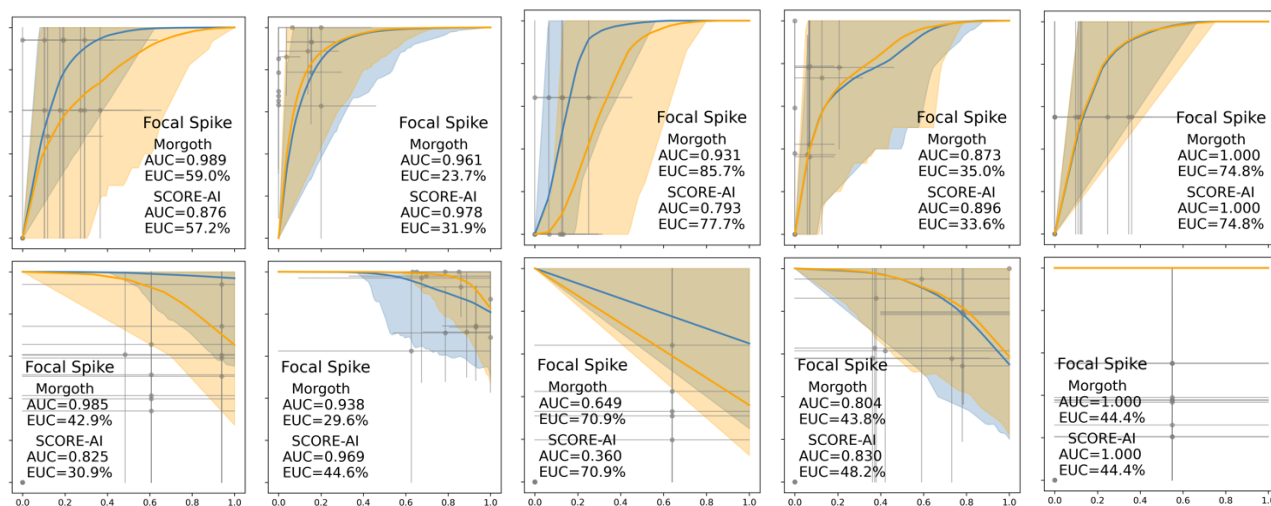
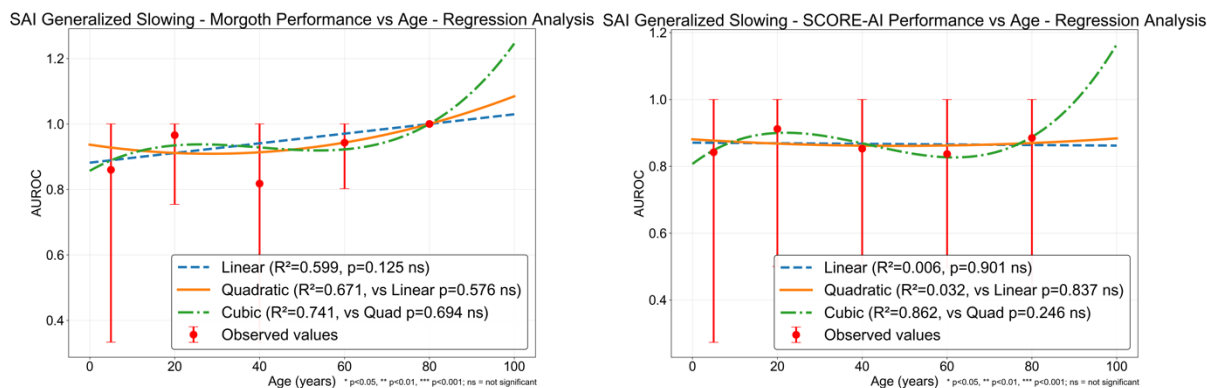


Norma vs Abnormal

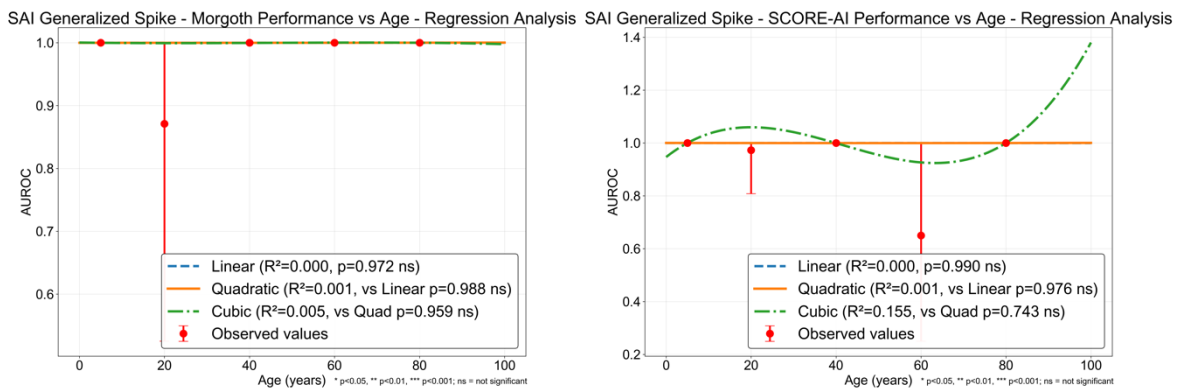
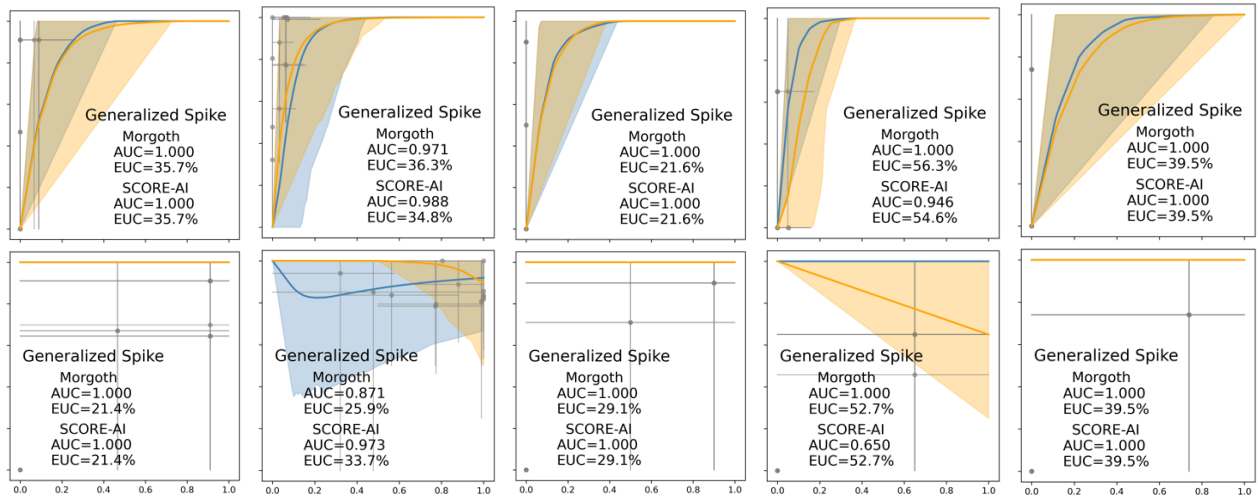


Focal Slowing Detection



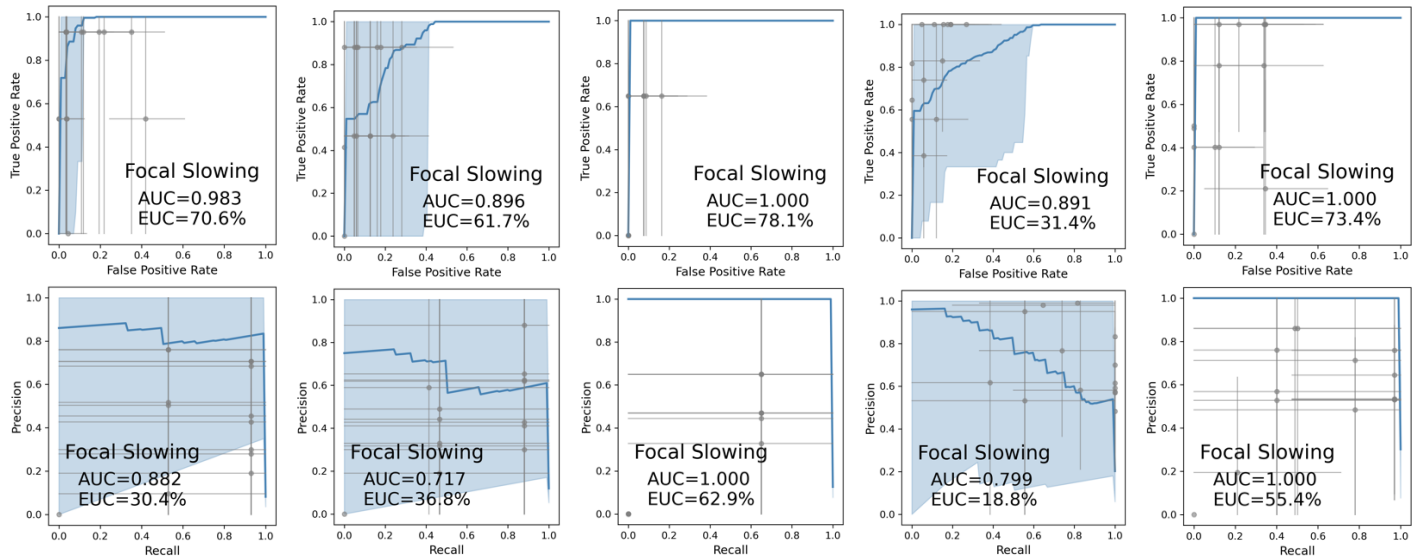


Focal Spike Detection

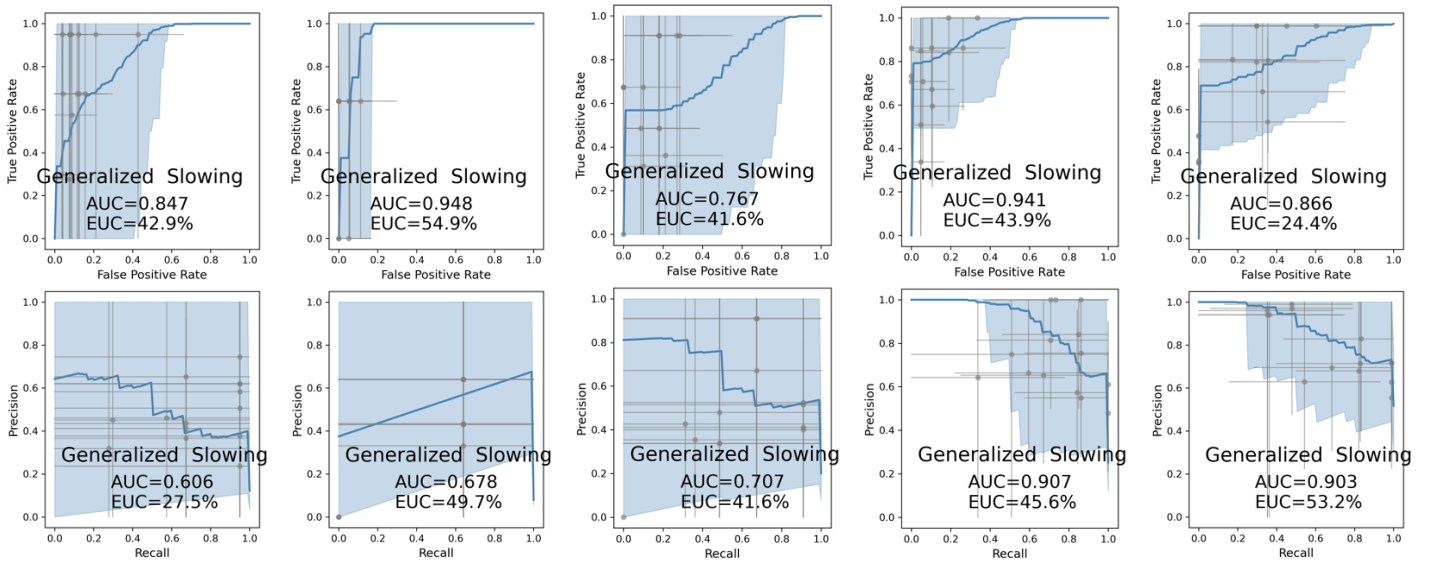


Generalized Spike Detection

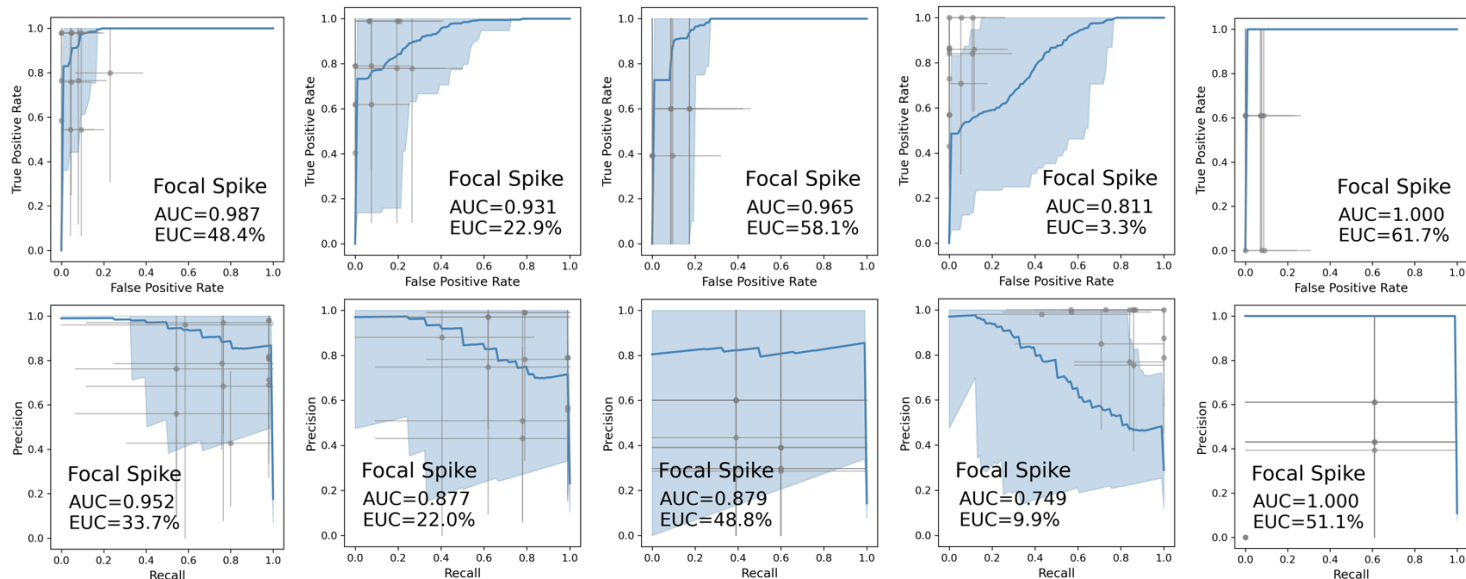
(D) External SAI dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue), SCORE-AI (yellow) and expert points with 95% CIs. The age groups from left to right are: [0, 10), [10, 30), [30, 50), [50, 70), and >70. Regression between age and AUC-RPC - if age significantly affect model performance. * $p<0.05$, ** $p<0.01$, *** $p<0.001$; ns = not significant.



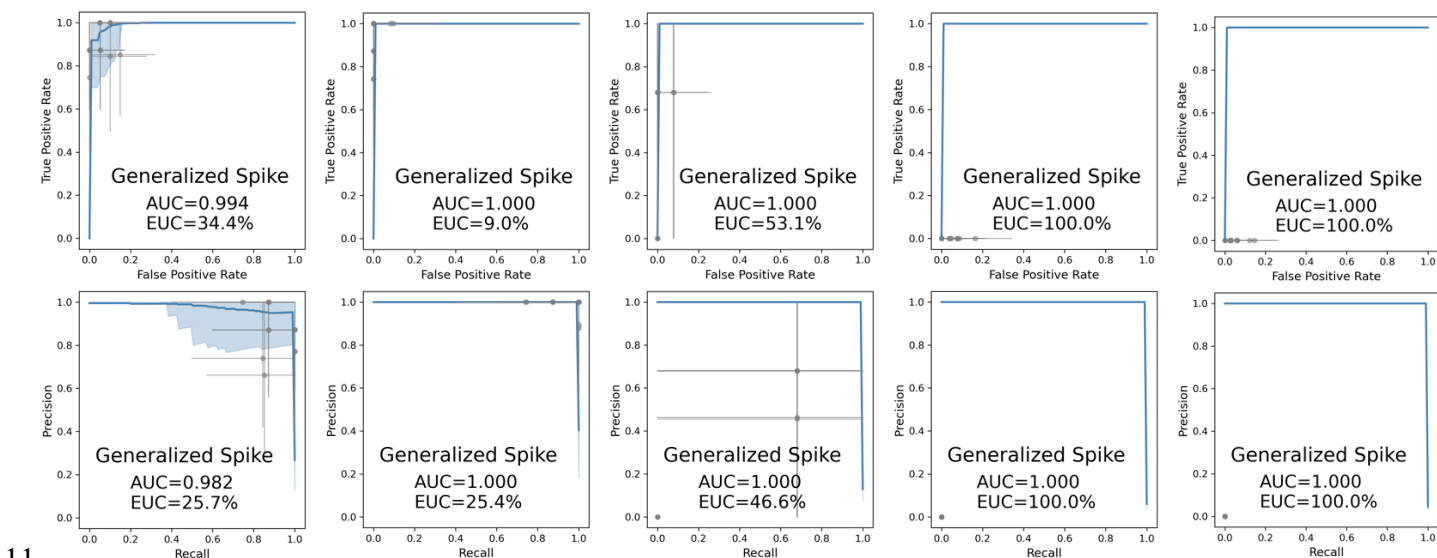
Focal Slowing Detection



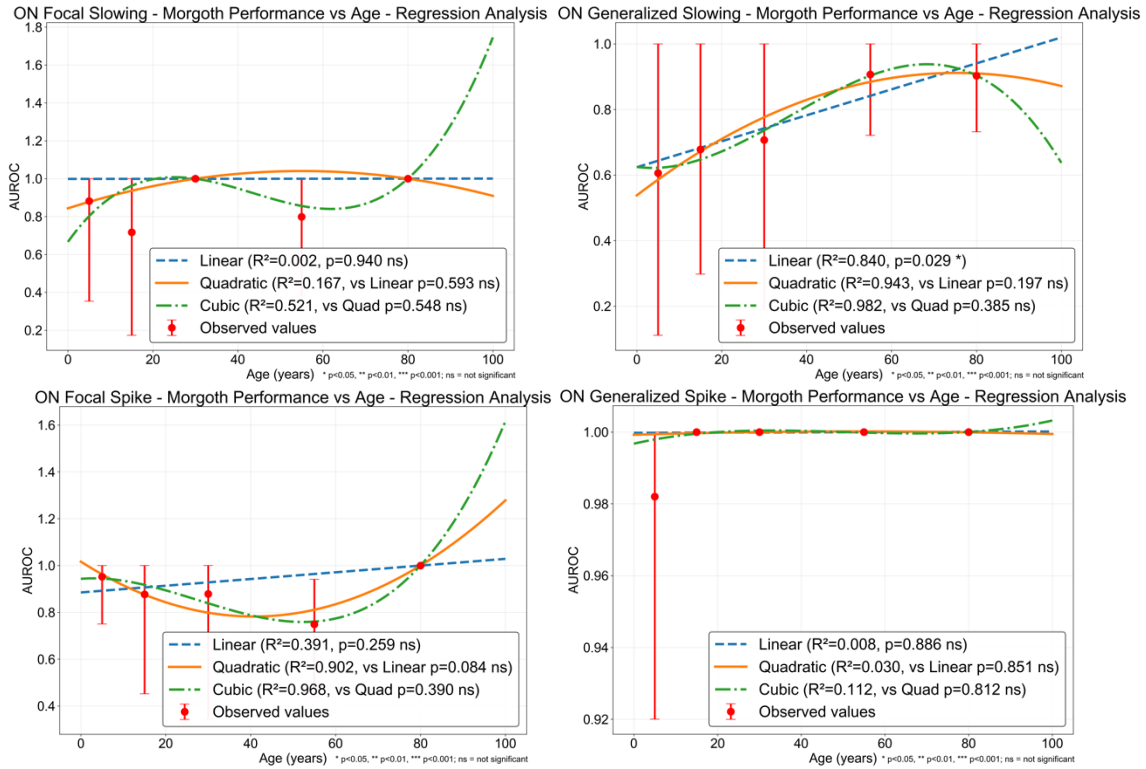
Generalized Slowing Detection



Focal Spike Detection



Generalized Spike Detection



(E) External ON dataset: Average ROC curves (odd rows) and PR curves (even rows) for Morgoth (blue) and expert points with 95% CIs. The age groups from left to right are: [0, 10), [10, 20), [20, 40), [40, 70), and >70. Regression between age and AUC-RPC - if age significantly affects model performance. * $p<0.05$, ** $p<0.01$, *** $p<0.001$; ns = not significant.

Figure S23. Age Group Specific Performance Results on Multi-expert Scored Datasets

Table S4. Comprehensive Results of Morgoth, Baselines, and Experts based on Gender Group

(A) Internal IIIC-Test dataset

Class	Model	AUC-ROC (F)	AUC-ROC (M)	EUC-ROC (F)	EUC-ROC (M)	AUC-PR (F)	AUC-PR (M)	EUC-PR (M)	EUC-PR (F)
Other	Morgoth	0.988 (0.985, 0.991)	0.985 (0.982, 0.989)	99.9% (100.0%, 100.0%)	98.8% (96.7%, 100.0%)	0.981 (0.976, 0.984)	0.985 (0.981, 0.988)	98.8% (96.7%, 100.0%)	99.5% (96.7%, 100.0%)
Other	Kaggle Winner	0.892 (0.882, 0.900)	0.936 (0.928, 0.943)	40.9% (33.3%, 51.7%)	80.2% (70.0%, 90.0%)	0.853 (0.840, 0.865)	0.924 (0.913, 0.936)	77.2% (64.9%, 86.7%)	26.2% (16.7%, 35.1%)
Other	SPaRCNet	0.777 (0.766, 0.788)	0.796 (0.781, 0.810)	1.2% (0.0%, 6.7%)	2.2% (0.0%, 3.3%)	0.617 (0.595, 0.640)	0.766 (0.745, 0.787)	1.6% (0.0%, 6.7%)	0.2% (0.0%, 3.3%)
Seizure	Morgoth	0.992 (0.990, 0.994)	0.976 (0.968, 0.983)	85.3% (76.7%, 90.0%)	67.9% (53.3%, 81.7%)	0.930 (0.908, 0.952)	0.702 (0.637, 0.779)	71.4% (56.7%, 90.0%)	92.3% (86.7%, 96.7%)
Seizure	Kaggle Winner	0.956 (0.950, 0.961)	0.962 (0.946, 0.972)	32.7% (23.3%, 43.3%)	54.1% (36.7%, 68.4%)	0.769 (0.744, 0.796)	0.714 (0.659, 0.758)	58.2% (43.2%, 75.1%)	69.0% (60.0%, 80.0%)
Seizure	SPaRCNet	0.846 (0.836, 0.857)	0.881 (0.866, 0.899)	0.3% (0.0%, 3.3%)	2.2% (0.0%, 6.7%)	0.334 (0.308, 0.360)	0.275 (0.221, 0.330)	6.5% (0.0%, 15.1%)	16.1% (8.2%, 26.7%)
LPD	Morgoth	0.989 (0.987, 0.991)	0.991 (0.987, 0.995)	100% (100%, 100%)	94.3% (86.7%, 100.0%)	0.982 (0.979, 0.985)	0.971 (0.964, 0.979)	93.4% (86.7%, 100.0%)	100% (100%, 100%)
LPD	Kaggle Winner	0.932 (0.926, 0.938)	0.967 (0.961, 0.973)	24.2% (16.7%, 33.3%)	63.8% (46.7%, 76.7%)	0.867 (0.855, 0.879)	0.881 (0.862, 0.898)	58.6% (44.9%, 73.3%)	42.0% (30.0%, 56.7%)
LPD	SPaRCNet	0.827 (0.816, 0.837)	0.877 (0.862, 0.891)	0.2% (0.0%, 3.3%)	7.2% (0.0%, 15.1%)	0.732 (0.711, 0.750)	0.655 (0.616, 0.698)	8.0% (0.0%, 16.7%)	0.6% (0.0%, 3.3%)

GPD	Morgoth	0.996 (0.995, 0.997)	0.990 (0.986, 0.993)	99.3% (96.7%, 100.0%)	99.3% (96.7%, 100.0%)	0.964 (0.951, 0.975)	0.955 (0.940, 0.966)	99.1% (94.9%, 100.0%)	99.3% (96.7%, 100.0%)
GPD	Kaggle Winner	0.983 (0.980, 0.986)	0.975 (0.968, 0.982)	93.1% (83.3%, 100.0%)	89.0% (78.2%, 98.4%)	0.860 (0.834, 0.886)	0.892 (0.868, 0.913)	83.6% (70.0%, 95.1%)	75.9% (60.0%, 86.7%)
GPD	SPaRCNet	0.939 (0.926, 0.951)	0.917 (0.902, 0.929)	48.5% (30.0%, 70.2%)	31.1% (16.7%, 43.3%)	0.745 (0.705, 0.778)	0.735 (0.703, 0.765)	28.5% (16.7%, 40.0%)	31.6% (18.2%, 50.0%)
LRDA	Morgoth	0.988 (0.983, 0.994)	0.992 (0.989, 0.996)	99.9% (96.7%, 100.0%)	97.5% (91.6%, 100.0%)	0.873 (0.831, 0.912)	0.942 (0.921, 0.962)	96.4% (90.0%, 100.0%)	99.9% (98.2%, 100.0%)
LRDA	Kaggle Winner	0.931 (0.918, 0.944)	0.948 (0.936, 0.958)	76.2% (60.0%, 93.3%)	56.0% (41.6%, 73.3%)	0.451 (0.388, 0.508)	0.655 (0.606, 0.707)	62.0% (43.3%, 76.7%)	73.7% (54.9%, 93.3%)
LRDA	SPaRCNet	0.835 (0.807, 0.861)	0.890 (0.869, 0.907)	18.3% (3.3%, 36.7%)	26.6% (14.9%, 43.3%)	0.223 (0.174, 0.271)	0.496 (0.434, 0.558)	33.9% (21.6%, 46.7%)	15.8% (1.6%, 30.0%)
GRDA	Morgoth	0.993 (0.990, 0.995)	0.986 (0.981, 0.990)	98.0% (93.3%, 100.0%)	97.2% (93.3%, 100.0%)	0.936 (0.913, 0.957)	0.941 (0.923, 0.957)	98.2% (93.3%, 100.0%)	97.5% (93.3%, 100.0%)
GRDA	Kaggle Winner	0.975 (0.970, 0.979)	0.947 (0.938, 0.955)	78.3% (66.7%, 90.0%)	51.0% (33.3%, 66.7%)	0.740 (0.698, 0.780)	0.710 (0.666, 0.752)	60.3% (40.0%, 78.4%)	67.0% (50.0%, 80.0%)
GRDA	SPaRCNet	0.929 (0.914, 0.941)	0.858 (0.842, 0.872)	30.9% (20.0%, 43.3%)	10.1% (6.7%, 16.7%)	0.560 (0.504, 0.611)	0.496 (0.439, 0.540)	13.8% (6.7%, 20.0%)	18.0% (10.0%, 30.0%)

1130

1131

1132 (B) Internal SN2-Test dataset

1133

Model	ROC AUC (F)	ROC AUC (M)	PR AUC (F)	PR AUC (M)	ROC EUC (F)	ROC EUC (M)	PR EUC (F)	PR EUC (M)
Morgoth	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	98.9% (91.7%, 100.0%)	100.0% (100.0%, 100.0%)	90.5% (83.3%, 95.8%)	93.0% (83.3%, 100.0%)
SpikeNet2	0.976 (0.959, 0.989)	0.978 (0.964, 0.990)	0.991 (0.983, 0.997)	0.992 (0.986, 0.997)	53.4% (33.3%, 75.0%)	65.5% (41.7%, 87.5%)	57.0% (37.5%, 79.2%)	74.4% (54.2%, 91.7%)

1134

1135 (C) Internal MoE dataset

1136

Task	ROC AUC (F)	ROC AUC (M)	PR AUC (F)	PR AUC (M)	ROC EUC (F)	ROC EUC (M)	PR EUC (F)	PR EUC (M)
IIIC - Seizure	0.937 (0.869, 0.980)	0.978 (0.958, 0.997)	0.685 (0.515, 0.838)	0.797 (0.595, 0.967)	56.0% (27.3%, 84.2%)	47.0% (27.3%, 75.1%)	86.2% (68.2%, 100.0%)	94.6% (79.4%, 100.0%)
IIIC - LPD	0.962 (0.916, 0.990)	0.963 (0.930, 0.990)	0.890 (0.801, 0.954)	0.813 (0.655, 0.921)	81.6% (65.8%, 95.5%)	87.3% (70.3%, 100.0%)	98.2% (93.1%, 100.0%)	95.0% (79.4%, 100.0%)
IIIC - GPD	0.971 (0.942, 0.994)	0.962 (0.919, 0.994)	0.787 (0.638, 0.912)	0.771 (0.586, 0.935)	57.0% (34.0%, 72.7%)	54.6% (24.9%, 84.2%)	82.9% (56.7%, 97.8%)	78.5% (50.0%, 100.0%)
IIIC - LRDA	0.996 (0.988, 1.000)	0.966 (0.924, 0.997)	0.881 (0.632, 1.000)	0.800 (0.623, 0.938)	79.8% (63.6%, 90.9%)	77.9% (59.1%, 97.8%)	93.8% (79.4%, 100.0%)	96.6% (90.9%, 100.0%)
IIIC - GRDA	nan (nan, nan)	0.994 (0.979, 1.000)	0.622 (0.011, 1.000)	0.897 (0.644, 1.000)	81.1% (65.8%, 100.0%)	57.3% (31.8%, 77.3%)	88.6% (59.1%, 100.0%)	88.9% (56.7%, 100.0%)
Focal Slowing Generalized Slowing	0.906 (0.874, 0.928)	0.931 (0.911, 0.955)	0.827 (0.783, 0.866)	0.854 (0.800, 0.898)	86.8% (77.3%, 95.5%)	88.5% (79.4%, 95.5%)	89.3% (81.8%, 95.5%)	90.8% (81.8%, 97.8%)
	0.870 (0.853, 0.894)	0.863 (0.837, 0.889)	0.839 (0.804, 0.867)	0.854 (0.818, 0.887)	94.0% (84.0%, 100.0%)	95.7% (86.4%, 100.0%)	94.4% (86.4%, 100.0%)	97.0% (86.4%, 100.0%)
Focal Spikes	0.959 (0.937, 0.974)	0.953 (0.927, 0.971)	0.849 (0.789, 0.891)	0.864 (0.793, 0.911)	80.7% (63.6%, 93.3%)	90.0% (79.4%, 100.0%)	84.8% (70.3%, 95.5%)	92.1% (86.4%, 100.0%)
Generalized Spikes	0.967 (0.948, 0.981)	0.964 (0.934, 0.988)	0.737 (0.614, 0.821)	0.798 (0.700, 0.889)	74.0% (54.5%, 90.9%)	76.6% (50.0%, 95.5%)	85.2% (65.8%, 100.0%)	86.5% (65.8%, 100.0%)
SPIKES	0.959 (0.944, 0.973)	0.947 (0.919, 0.973)	0.862 (0.818, 0.904)	0.862 (0.815, 0.921)	85.4% (68.2%, 97.8%)	86.5% (72.7%, 95.5%)	86.5% (70.3%, 97.8%)	88.7% (77.3%, 100.0%)
Burst Suppression	0.999 (0.997, 1.000)	0.999 (0.998, 1.000)	0.954 (0.877, 0.999)	0.965 (0.911, 0.996)	67.9% (50.0%, 81.8%)	85.0% (72.7%, 97.8%)	95.0% (81.8%, 100.0%)	98.1% (90.9%, 100.0%)
Awake	0.991 (0.983, 0.996)	0.982 (0.963, 0.996)	0.997 (0.994, 0.999)	0.993 (0.985, 0.998)	93.3% (81.8%, 100.0%)	93.5% (81.8%, 100.0%)	96.0% (86.4%, 100.0%)	94.2% (86.4%, 100.0%)
N1	0.684 (0.612, 0.762)	0.744 (0.596, 0.852)	0.127 (0.078, 0.182)	0.215 (0.071, 0.415)	11.9% (9.1%, 18.2%)	17.9% (9.1%, 29.7%)	12.0% (9.1%, 18.2%)	17.6% (9.1%, 27.3%)
N2	0.986 (0.973, 0.994)	0.979 (0.964, 0.992)	0.926 (0.851, 0.976)	0.917 (0.857, 0.964)	55.9% (22.7%, 72.7%)	51.8% (22.7%, 77.3%)	73.0% (43.1%, 100.0%)	60.1% (34.0%, 84.2%)

1137

1138 (D) External SAI dataset

1139

Task	Model	ROC AUC (F)	ROC AUC (M)	PR AUC (F)	PR AUC (M)	ROC EUC (F)	ROC EUC (M)	PR EUC (F)	PR EUC (M)
Generalized Spike	Morgoth	1.000 (1.000, 1.000)	0.975 (0.822, 1.000)	1.000 (1.000, 1.000)	0.933 (0.726, 1.000)	40.5% (36.1%, 42.9%)	35.7% (15.0%, 50.0%)	40.5% (35.7%, 49.3%)	22.9% (14.3%, 40.7%)
Generalized Spike	SCORE-AI	0.997 (0.991, 1.000)	0.982 (0.889, 1.000)	0.977 (0.917, 1.000)	0.950 (0.667, 1.000)	26.2% (14.6%, 41.8%)	25.7% (15.0%, 47.1%)	28.6% (21.8%, 35.4%)	25.7% (14.3%, 42.1%)
Focal Spike	Morgoth	0.968 (0.816, 1.000)	0.918 (0.686, 1.000)	0.903 (0.659, 1.000)	0.889 (0.630, 1.000)	48.1% (23.9%, 76.1%)	46.4% (17.7%, 75.2%)	46.2% (28.6%, 76.1%)	47.1% (13.9%, 68.0%)
Focal Spike	SCORE-AI	0.916 (0.799, 1.000)	0.954 (0.816, 1.000)	0.786 (0.500, 1.000)	0.865 (0.433, 1.000)	26.2% (14.3%, 47.5%)	52.1% (21.4%, 82.3%)	24.3% (14.3%, 40.4%)	47.1% (24.8%, 71.4%)
Focal Slowing	Morgoth	0.974 (0.833, 1.000)	0.990 (0.918, 1.000)	0.961 (0.679, 1.000)	0.988 (0.900, 1.000)	54.3% (42.9%, 64.3%)	62.1% (42.9%, 77.0%)	49.3% (28.6%, 64.3%)	61.4% (30.2%, 84.1%)

Focal Slowing	SCORE-AI	0.827 (0.562, 0.988)	0.959 (0.898, 1.000)	0.731 (0.313, 0.959)	0.919 (0.772, 1.000)	17.5% (0.0%, 35.7%)	41.4% (21.4%, 69.8%)	22.1% (3.4%, 35.7%)	39.3% (15.9%, 57.1%)
Generalized Slowing	Morgoth	0.961 (0.838, 1.000)	0.954 (0.852, 1.000)	0.921 (0.759, 1.000)	0.930 (0.802, 1.000)	45.0% (28.6%, 73.8%)	37.9% (21.4%, 57.1%)	45.0% (30.2%, 57.1%)	36.4% (28.6%, 48.4%)
Generalized Slowing	SCORE-AI	0.932 (0.840, 1.000)	0.963 (0.889, 1.000)	0.772 (0.568, 1.000)	0.913 (0.602, 1.000)	35.0% (15.9%, 55.5%)	42.1% (28.6%, 66.6%)	29.3% (15.9%, 42.9%)	33.6% (23.0%, 42.9%)
Normal	Morgoth	0.995 (0.950, 1.000)	0.971 (0.874, 1.000)	0.997 (0.967, 1.000)	0.930 (0.726, 1.000)	49.1% (28.6%, 64.3%)	22.6% (3.4%, 35.7%)	59.8% (42.9%, 78.6%)	37.4% (7.1%, 75.2%)
Normal	SCORE-AI	0.991 (0.919, 1.000)	1.000 (1.000, 1.000)	0.993 (0.939, 1.000)	1.000 (1.000, 1.000)	45.8% (21.4%, 68.0%)	29.4% (14.3%, 35.7%)	52.7% (21.4%, 75.2%)	61.7% (35.7%, 78.6%)

(E) External ON dataset

Task	ROC AUC (F)	ROC AUC (M)	PR AUC (F)	PR AUC (M)	ROC EUC (F)	ROC EUC (M)	PR EUC (F)	PR EUC (M)
Focal Slowing	0.998 (0.987, 1.000)	0.873 (0.689, 0.995)	0.987 (0.910, 1.000)	0.742 (0.382, 0.975)	82.0% (67.7%, 89.5%)	38.3% (7.1%, 78.6%)	47.9% (10.5%, 57.1%)	31.6% (0.0%, 71.4%)
Generalized Slowing	0.847 (0.714, 0.950)	0.966 (0.916, 1.000)	0.698 (0.430, 0.878)	0.878 (0.687, 1.000)	24.7% (0.0%, 71.4%)	56.8% (14.3%, 92.9%)	24.1% (0.0%, 71.4%)	45.6% (7.1%, 78.6%)
Focal Spike	0.874 (0.737, 0.970)	1.000 (1.000, 1.000)	0.754 (0.500, 0.931)	1.000 (1.000, 1.000)	11.5% (0.0%, 42.9%)	63.6% (42.7%, 78.6%)	15.9% (0.0%, 50.0%)	46.5% (14.3%, 57.1%)
Generalized Spike	0.999 (0.994, 1.000)	0.995 (0.978, 1.000)	0.995 (0.968, 1.000)	0.970 (0.869, 1.000)	47.1% (21.4%, 57.1%)	39.8% (14.3%, 57.1%)	31.0% (7.1%, 57.1%)	32.4% (7.1%, 64.3%)

Table S5 Significance Test of Gender Group Results

(A) Internal IIIC-Test dataset

Class	Model	Metric	Mean F	Mean M	Z	p-value
Other	Morgoth	AUC-ROC	0.988	0.985	1.276	0.2021
Other	Morgoth	EUC-ROC	99.9	98.8	1.307	0.1913
Other	Morgoth	AUC-PR	0.981	0.985	-1.475	0.1402
Other	Morgoth	EUC-PR	99.5	98.8	0.588	0.5566
Other	Kaggle Winner	AUC-ROC	0.892	0.936	-7.361	0
Other	Kaggle Winner	EUC-ROC	40.9	80.2	-5.669	0
Other	Kaggle Winner	AUC-PR	0.853	0.924	-8.193	0
Other	Kaggle Winner	EUC-PR	26.2	77.2	-7.008	0
Other	SPaRCNet	AUC-ROC	0.777	0.796	-2.046	0.0407
Other	SPaRCNet	EUC-ROC	1.2	2.2	-0.525	0.5997
Other	SPaRCNet	AUC-PR	0.617	0.766	-9.489	0
Other	SPaRCNet	EUC-PR	0.2	1.6	-0.735	0.4625
Seizure	Morgoth	AUC-ROC	0.992	0.976	4.04	0.0001
Seizure	Morgoth	EUC-ROC	85.3	67.9	2.175	0.0296
Seizure	Morgoth	AUC-PR	0.93	0.702	6.012	0
Seizure	Morgoth	EUC-PR	92.3	71.4	2.356	0.0185
Seizure	Kaggle Winner	AUC-ROC	0.956	0.962	-0.833	0.4048
Seizure	Kaggle Winner	EUC-ROC	32.7	54.1	-2.238	0.0252
Seizure	Kaggle Winner	AUC-PR	0.769	0.714	1.928	0.0539
Seizure	Kaggle Winner	EUC-PR	69	58.2	1.124	0.2608
Seizure	SPaRCNet	AUC-ROC	0.846	0.881	-3.508	0.0005
Seizure	SPaRCNet	EUC-ROC	0.3	2.2	-0.997	0.3186
Seizure	SPaRCNet	AUC-PR	0.334	0.275	1.915	0.0555
Seizure	SPaRCNet	EUC-PR	16.1	6.5	1.576	0.1151
LPD	Morgoth	AUC-ROC	0.989	0.991	-0.877	0.3807
LPD	Morgoth	EUC-ROC	100	94.3	1.68	0.093
LPD	Morgoth	AUC-PR	0.982	0.971	2.669	0.0076

LPD	Morgoth	EUC-PR	100	93.4	1.945	0.0517
LPD	Kaggle Winner	AUC-ROC	0.932	0.967	-8.085	0
LPD	Kaggle Winner	EUC-ROC	24.2	63.8	-4.528	0
LPD	Kaggle Winner	AUC-PR	0.867	0.881	-1.268	0.2046
LPD	Kaggle Winner	EUC-PR	42	58.6	-1.669	0.095
LPD	SPaRCNet	AUC-ROC	0.827	0.877	-5.474	0
LPD	SPaRCNet	EUC-ROC	0.2	7.2	-1.775	0.0758
LPD	SPaRCNet	AUC-PR	0.732	0.655	3.324	0.0009
LPD	SPaRCNet	EUC-PR	0.6	8	-1.704	0.0884
GPD	Morgoth	AUC-ROC	0.996	0.99	3.231	0.0012
GPD	Morgoth	EUC-ROC	99.3	99.3	0	1
GPD	Morgoth	AUC-PR	0.964	0.955	0.997	0.3187
GPD	Morgoth	EUC-PR	99.3	99.1	0.129	0.8973
GPD	Kaggle Winner	AUC-ROC	0.983	0.975	2.059	0.0395
GPD	Kaggle Winner	EUC-ROC	93.1	89	0.613	0.5397
GPD	Kaggle Winner	AUC-PR	0.86	0.892	-1.824	0.0681
GPD	Kaggle Winner	EUC-PR	75.9	83.6	-0.824	0.4101
GPD	SPaRCNet	AUC-ROC	0.939	0.917	2.344	0.0191
GPD	SPaRCNet	EUC-ROC	48.5	31.1	1.415	0.1571
GPD	SPaRCNet	AUC-PR	0.745	0.735	0.409	0.6823
GPD	SPaRCNet	EUC-PR	31.6	28.5	0.308	0.7579
LRDA	Morgoth	AUC-ROC	0.988	0.992	-1.203	0.2291
LRDA	Morgoth	EUC-ROC	99.9	97.5	1.042	0.2972
LRDA	Morgoth	AUC-PR	0.873	0.942	-2.979	0.0029
LRDA	Morgoth	EUC-PR	99.9	96.4	1.35	0.1769
LRDA	Kaggle Winner	AUC-ROC	0.931	0.948	-1.957	0.0504
LRDA	Kaggle Winner	EUC-ROC	76.2	56	1.722	0.085
LRDA	Kaggle Winner	AUC-PR	0.451	0.655	-5.098	0
LRDA	Kaggle Winner	EUC-PR	73.7	62	0.901	0.3675
LRDA	SPaRCNet	AUC-ROC	0.835	0.89	-3.265	0.0011
LRDA	SPaRCNet	EUC-ROC	18.3	26.6	-0.742	0.458
LRDA	SPaRCNet	AUC-PR	0.223	0.496	-6.798	0
LRDA	SPaRCNet	EUC-PR	15.8	33.9	-1.872	0.0612
GRDA	Morgoth	AUC-ROC	0.993	0.986	2.665	0.0077
GRDA	Morgoth	EUC-ROC	98	97.2	0.331	0.7407
GRDA	Morgoth	AUC-PR	0.936	0.941	-0.352	0.7245
GRDA	Morgoth	EUC-PR	97.5	98.2	-0.29	0.7721
GRDA	Kaggle Winner	AUC-ROC	0.975	0.947	5.706	0
GRDA	Kaggle Winner	EUC-ROC	78.3	51	2.628	0.0086
GRDA	Kaggle Winner	AUC-PR	0.74	0.71	0.99	0.3223
GRDA	Kaggle Winner	EUC-PR	67	60.3	0.539	0.5899
GRDA	SPaRCNet	AUC-ROC	0.929	0.858	6.896	0
GRDA	SPaRCNet	EUC-ROC	30.9	10.1	3.216	0.0013
GRDA	SPaRCNet	AUC-PR	0.56	0.496	1.705	0.0882
GRDA	SPaRCNet	EUC-PR	18	13.8	0.685	0.493

(B) Internal SN2-Test dataset

Model	Metric	Mean F	Mean M	Z	p-value
Morgoth	ROC AUC	1	1		

Morgoth	PR AUC	1	1		
Morgoth	ROC EUC	98.9	100	-0.52	0.6034
Morgoth	PR EUC	90.5	93	-0.47	0.6385
SpikeNet2	ROC AUC	0.976	0.978	-0.197	0.8434
SpikeNet2	PR AUC	0.991	0.992	-0.22	0.8257
SpikeNet2	ROC EUC	53.4	65.5	-0.766	0.4438
SpikeNet2	PR EUC	57	74.4	-1.216	0.2239

(C) Internal MoE dataset

Task	Metric	Mean F	Mean M	Z	p-value
IIIC - Seizure	ROC AUC	0.937	0.978	-1.366	0.1719
IIIC - Seizure	PR AUC	0.685	0.797	-0.891	0.3728
IIIC - Seizure	ROC EUC	56	47	0.475	0.635
IIIC - Seizure	PR EUC	86.2	94.6	-0.869	0.3848
IIIC - LPD	ROC AUC	0.962	0.963	-0.041	0.9672
IIIC - LPD	PR AUC	0.89	0.813	0.984	0.3253
IIIC - LPD	ROC EUC	81.6	87.3	-0.532	0.5947
IIIC - LPD	PR EUC	98.2	95	0.577	0.5637
IIIC - GPD	ROC AUC	0.971	0.962	0.387	0.6991
IIIC - GPD	PR AUC	0.787	0.771	0.141	0.8876
IIIC - GPD	ROC EUC	57	54.6	0.133	0.8943
IIIC - GPD	PR EUC	82.9	78.5	0.266	0.7899
IIIC - LRDA	ROC AUC	0.996	0.966	1.59	0.1119
IIIC - LRDA	PR AUC	0.881	0.8	0.655	0.5122
IIIC - LRDA	ROC EUC	79.8	77.9	0.157	0.875
IIIC - LRDA	PR EUC	93.8	96.6	-0.487	0.626
IIIC - GRDA	ROC AUC	nan	nan	nan	nan
IIIC - GRDA	PR AUC	0.622	0.897	-1.026	0.3051
IIIC - GRDA	ROC EUC	81.1	57.3	1.639	0.1012
IIIC - GRDA	PR EUC	88.6	88.9	-0.02	0.9842
Focal Slowing	ROC AUC	0.906	0.931	-1.407	0.1595
Focal Slowing	PR AUC	0.827	0.854	-0.824	0.4099
Focal Slowing	ROC EUC	86.8	88.5	-0.274	0.7839
Focal Slowing	PR EUC	89.3	90.8	-0.279	0.7801
Generalized Slowing	ROC AUC	0.87	0.863	0.414	0.6786
Generalized Slowing	PR AUC	0.839	0.854	-0.629	0.5291
Generalized Slowing	ROC EUC	94	95.7	-0.317	0.751
Generalized Slowing	PR EUC	94.4	97	-0.53	0.5962
Focal Spikes	ROC AUC	0.959	0.953	0.409	0.6825
Focal Spikes	PR AUC	0.849	0.864	-0.377	0.7062
Focal Spikes	ROC EUC	80.7	90	-1.009	0.3132
Focal Spikes	PR EUC	84.8	92.1	-0.999	0.3176
Generalized Spikes	ROC AUC	0.967	0.964	0.186	0.8526
Generalized Spikes	PR AUC	0.737	0.798	-0.853	0.3936
Generalized Spikes	ROC EUC	74	76.6	-0.175	0.8611
Generalized Spikes	PR EUC	85.2	86.5	-0.105	0.9161
SPIKES	ROC AUC	0.959	0.947	0.767	0.4428
SPIKES	PR AUC	0.862	0.862	0	1
SPIKES	ROC EUC	85.4	86.5	-0.115	0.9081
SPIKES	PR EUC	86.5	88.7	-0.242	0.8089
Burst Suppression	ROC AUC	0.999	0.999	0	1
Burst Suppression	PR AUC	0.954	0.965	-0.29	0.7718
Burst Suppression	ROC EUC	67.9	85	-1.655	0.098

Burst Suppression	PR EUC	95	98.1	-0.597	0.5504
Awake	ROC AUC	0.991	0.982	0.995	0.3199
Awake	PR AUC	0.997	0.993	1.126	0.2603
Awake	ROC EUC	93.3	93.5	-0.03	0.9757
Awake	PR EUC	96	94.2	0.367	0.7137
N1	ROC AUC	0.684	0.744	-0.793	0.428
N1	PR AUC	0.127	0.215	-0.96	0.3371
N1	ROC EUC	11.9	17.9	-1.044	0.2963
N1	PR EUC	12	17.6	-1.079	0.2807
N2	ROC AUC	0.986	0.979	0.784	0.433
N2	PR AUC	0.926	0.917	0.214	0.8302
N2	ROC EUC	55.9	51.8	0.217	0.8281
N2	PR EUC	73	60.1	0.666	0.5051

(D) External SAI dataset

Task	Model	Metric	Mean F	Mean M	Z	p-value
Generalized Spike	Morgoth	ROC AUC	1	0.975	0.551	0.5819
Generalized Spike	Morgoth	PR AUC	1	0.933	0.959	0.3378
Generalized Spike	Morgoth	ROC EUC	40.5	35.7	0.528	0.5977
Generalized Spike	Morgoth	PR EUC	40.5	22.9	2.323	0.0202
Generalized Spike	SCORE-AI	ROC AUC	0.997	0.982	0.528	0.5975
Generalized Spike	SCORE-AI	PR AUC	0.977	0.95	0.308	0.7578
Generalized Spike	SCORE-AI	ROC EUC	26.2	25.7	0.047	0.9628
Generalized Spike	SCORE-AI	PR EUC	28.6	25.7	0.367	0.7134
Focal Spike	Morgoth	ROC AUC	0.968	0.918	0.539	0.5902
Focal Spike	Morgoth	PR AUC	0.903	0.889	0.109	0.9131
Focal Spike	Morgoth	ROC EUC	48.1	46.4	0.086	0.9316
Focal Spike	Morgoth	PR EUC	46.2	47.1	-0.049	0.9609
Focal Spike	SCORE-AI	ROC AUC	0.916	0.954	-0.547	0.5846
Focal Spike	SCORE-AI	PR AUC	0.786	0.865	-0.41	0.6821
Focal Spike	SCORE-AI	ROC EUC	26.2	52.1	-1.464	0.1433
Focal Spike	SCORE-AI	PR EUC	34.3	47.1	-1.673	0.1943
Focal Slowing	Morgoth	ROC AUC	0.974	0.99	-0.337	0.736
Focal Slowing	Morgoth	PR AUC	0.961	0.988	-0.315	0.7529
Focal Slowing	Morgoth	ROC EUC	54.3	62.1	-0.759	0.4476
Focal Slowing	Morgoth	PR EUC	49.3	61.4	-0.734	0.4632
Focal Slowing	SCORE-AI	ROC AUC	0.827	0.959	-1.181	0.2375
Focal Slowing	SCORE-AI	PR AUC	0.731	0.919	-1.076	0.282
Focal Slowing	SCORE-AI	ROC EUC	17.5	41.4	-1.558	0.1193
Focal Slowing	SCORE-AI	PR EUC	22.1	39.3	-1.288	0.1978
Generalized Slowing	Morgoth	ROC AUC	0.961	0.954	0.125	0.9005
Generalized Slowing	Morgoth	PR AUC	0.921	0.93	-0.113	0.9099
Generalized Slowing	Morgoth	ROC EUC	45	37.9	0.483	0.6289
Generalized Slowing	Morgoth	PR EUC	45	36.4	1.009	0.3128
Generalized Slowing	SCORE-AI	ROC AUC	0.932	0.963	-0.624	0.5326
Generalized Slowing	SCORE-AI	PR AUC	0.772	0.913	-0.941	0.3467
Generalized Slowing	SCORE-AI	ROC EUC	35	42.1	-0.507	0.6121
Generalized Slowing	SCORE-AI	PR EUC	29.3	33.6	-0.503	0.6153
Normal	Morgoth	ROC AUC	0.995	0.971	0.694	0.4877
Normal	Morgoth	PR AUC	0.997	0.93	0.952	0.3413
Normal	Morgoth	ROC EUC	49.1	42.6	1.158	0.1309
Normal	Morgoth	PR EUC	59.8	37.4	1.142	0.2535
Normal	SCORE-AI	ROC AUC	0.991	1	-0.436	0.6632
Normal	SCORE-AI	PR AUC	0.993	1	-0.45	0.6528
Normal	SCORE-AI	ROC EUC	45.8	29.4	1.254	0.21

Normal	SCORE-AI	PR EUC	52.7	61.7	-0.513	0.6082
--------	----------	--------	------	------	--------	--------

(E) External ON dataset

Task	Metric	Mean F	Mean M	Z	p-value
Focal Slowing	ROC AUC	0.998	0.873	1.6	0.1096
Focal Slowing	PR AUC	0.987	0.742	1.601	0.1093
Focal Slowing	ROC EUC	82	38.3	2.292	0.0219
Focal Slowing	PR EUC	47.9	31.6	0.749	0.4536
Generalized Slowing	ROC AUC	0.847	0.966	-1.862	0.0626
Generalized Slowing	PR AUC	0.698	0.878	-1.291	0.1967
Generalized Slowing	ROC EUC	24.7	56.8	-1.185	0.236
Generalized Slowing	PR EUC	24.1	45.6	-0.834	0.4042
Focal Spike	ROC AUC	0.874	1	-2.12	0.034
Focal Spike	PR AUC	0.754	1	-2.237	0.0253
Focal Spike	ROC EUC	11.5	63.6	-3.651	0.0003
Focal Spike	PR EUC	15.9	46.5	-1.823	0.0684
Generalized Spike	ROC AUC	0.999	0.995	0.688	0.4917
Generalized Spike	PR AUC	0.995	0.97	0.727	0.4674
Generalized Spike	ROC EUC	47.1	39.8	0.513	0.6076
Generalized Spike	PR EUC	31	32.4	-0.072	0.9424

Table S7 Comprehensive Results of Morgoth, Baselines, and Experts based on Age Group

(A) Internal IIIC-Test dataset

Model	Class	Age Group	AUC-ROC	AUC-PR	EUC-ROC	EUC-PR
Morgoth	GPD	IIIC 1-10	0.922 (0.856, 0.961)	0.871 (0.776, 0.930)	47.4% (20.3%, 70.4%)	46.7% (18.5%, 70.4%)
Morgoth	GPD	IIIC 10-20	0.982 (0.958, 0.996)	0.907 (0.796, 0.975)	71.9% (51.6%, 93.3%)	67.6% (43.3%, 90.0%)
Morgoth	GPD	IIIC 20-30	0.987 (0.979, 0.994)	0.943 (0.907, 0.971)	88.3% (74.9%, 96.7%)	86.6% (73.3%, 96.7%)
Morgoth	GPD	IIIC 30-40	0.997 (0.994, 0.999)	0.994 (0.986, 0.998)	99.1% (96.7%, 100.0%)	98.9% (96.7%, 100.0%)
Morgoth	GPD	IIIC 40-50	0.996 (0.991, 1.000)	0.965 (0.914, 1.000)	92.6% (83.3%, 100.0%)	88.7% (74.9%, 96.7%)
Morgoth	GPD	IIIC 50-60	0.997 (0.993, 1.000)	0.989 (0.979, 0.997)	99.7% (96.7%, 100.0%)	99.7% (96.7%, 100.0%)
Morgoth	GPD	IIIC 60-70	0.997 (0.994, 0.998)	0.939 (0.901, 0.971)	95.9% (88.2%, 100.0%)	92.2% (83.3%, 100.0%)
Morgoth	GPD	IIIC 70-80	0.994 (0.991, 0.996)	0.938 (0.908, 0.964)	92.7% (83.3%, 100.0%)	91.1% (83.3%, 98.4%)
Morgoth	GPD	IIIC 80-90	0.998 (0.995, 0.999)	0.982 (0.968, 0.994)	95.5% (86.7%, 100.0%)	89.0% (81.6%, 93.3%)
Morgoth	GPD	IIIC >90	0.994 (0.984, 1.000)	0.912 (0.791, 1.000)	83.3% (70.0%, 100.0%)	64.7% (53.3%, 75.1%)
Morgoth	GRDA	IIIC 1-10	0.980 (0.951, 1.000)	0.781 (0.375, 1.000)	76.6% (63.0%, 100.0%)	63.7% (46.2%, 79.7%)
Morgoth	GRDA	IIIC 10-20	0.986 (0.974, 0.997)	0.943 (0.878, 0.986)	84.1% (70.0%, 96.7%)	71.9% (58.2%, 83.3%)
Morgoth	GRDA	IIIC 20-30	0.979 (0.965, 0.991)	0.943 (0.896, 0.978)	85.9% (73.3%, 96.7%)	86.1% (74.9%, 96.7%)
Morgoth	GRDA	IIIC 30-40	0.994 (0.988, 0.998)	0.984 (0.971, 0.994)	92.2% (83.3%, 100.0%)	89.5% (80.0%, 96.7%)
Morgoth	GRDA	IIIC 40-50	0.987 (0.977, 0.995)	0.949 (0.913, 0.979)	85.5% (73.3%, 96.7%)	84.8% (73.3%, 96.7%)
Morgoth	GRDA	IIIC 50-60	0.987 (0.979, 0.995)	0.871 (0.783, 0.945)	91.7% (80.0%, 96.7%)	87.1% (71.6%, 95.1%)
Morgoth	GRDA	IIIC 60-70	0.995 (0.991, 0.998)	0.944 (0.912, 0.973)	93.4% (86.7%, 100.0%)	86.9% (80.0%, 93.3%)

Morgoth	GRDA	IIIC 70-80	0.990 (0.984, 0.994)	0.929 (0.894, 0.958)	92.3% (83.3%, 100.0%)	88.2% (80.0%, 95.1%)
Morgoth	GRDA	IIIC 80-90	0.994 (0.988, 0.999)	0.900 (0.814, 0.963)	89.2% (76.7%, 98.4%)	85.6% (70.0%, 96.7%)
Morgoth	GRDA	IIIC >90	0.990 (0.973, 1.000)	0.923 (0.794, 1.000)	93.1% (83.3%, 100.0%)	88.7% (76.7%, 96.7%)
Morgoth	LPD	IIIC 0-1	0.970 (0.908, 1.000)	0.896 (0.638, 1.000)	75.5% (38.1%, 100.0%)	58.1% (21.3%, 85.7%)
Morgoth	LPD	IIIC 1-10	0.906 (0.754, 0.992)	0.803 (0.630, 0.940)	47.3% (27.7%, 81.5%)	43.5% (25.9%, 64.9%)
Morgoth	LPD	IIIC 10-20	0.981 (0.968, 0.993)	0.963 (0.933, 0.984)	82.3% (66.7%, 93.3%)	79.2% (61.6%, 90.0%)
Morgoth	LPD	IIIC 20-30	0.993 (0.986, 0.999)	0.975 (0.945, 0.993)	87.5% (78.2%, 93.3%)	87.6% (80.0%, 93.3%)
Morgoth	LPD	IIIC 30-40	0.980 (0.966, 0.994)	0.648 (0.428, 0.852)	65.4% (50.0%, 83.3%)	64.3% (48.2%, 80.0%)
Morgoth	LPD	IIIC 40-50	0.994 (0.986, 0.999)	0.989 (0.975, 0.998)	90.7% (80.0%, 96.7%)	90.1% (80.0%, 96.7%)
Morgoth	LPD	IIIC 50-60	0.996 (0.994, 0.998)	0.990 (0.982, 0.995)	98.2% (93.3%, 100.0%)	98.3% (93.3%, 100.0%)
Morgoth	LPD	IIIC 60-70	0.987 (0.983, 0.991)	0.977 (0.970, 0.983)	94.7% (86.7%, 100.0%)	96.0% (90.0%, 100.0%)
Morgoth	LPD	IIIC 70-80	0.993 (0.990, 0.996)	0.983 (0.975, 0.989)	95.8% (90.0%, 100.0%)	96.2% (90.0%, 100.0%)
Morgoth	LPD	IIIC 80-90	0.991 (0.988, 0.994)	0.988 (0.983, 0.992)	96.6% (91.6%, 100.0%)	97.3% (93.3%, 100.0%)
Morgoth	LPD	IIIC >90	0.998 (0.994, 1.000)	0.996 (0.987, 1.000)	81.9% (61.6%, 100.0%)	67.6% (56.7%, 80.0%)
Morgoth	LRDA	IIIC 1-10	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	97.8% (69.4%, 100.0%)
Morgoth	LRDA	IIIC 10-20	0.989 (0.975, 1.000)	0.426 (0.143, 1.000)	99.6% (95.0%, 100.0%)	99.1% (81.7%, 100.0%)
Morgoth	LRDA	IIIC 20-30	0.988 (0.979, 0.995)	0.919 (0.870, 0.966)	92.2% (83.3%, 100.0%)	92.7% (84.9%, 100.0%)
Morgoth	LRDA	IIIC 30-40	0.999 (0.997, 1.000)	0.938 (0.715, 1.000)	96.4% (84.8%, 100.0%)	56.3% (44.8%, 100.0%)
Morgoth	LRDA	IIIC 40-50	0.990 (0.984, 0.996)	0.892 (0.781, 0.956)	84.7% (71.6%, 96.7%)	76.1% (63.3%, 86.7%)
Morgoth	LRDA	IIIC 50-60	0.997 (0.993, 0.999)	0.967 (0.940, 0.987)	96.7% (93.3%, 100.0%)	95.5% (90.0%, 100.0%)
Morgoth	LRDA	IIIC 60-70	0.984 (0.968, 0.996)	0.917 (0.877, 0.952)	94.0% (86.7%, 100.0%)	93.8% (86.7%, 100.0%)
Morgoth	LRDA	IIIC 70-80	0.992 (0.987, 0.995)	0.930 (0.895, 0.957)	95.7% (88.2%, 100.0%)	96.1% (88.2%, 100.0%)
Morgoth	LRDA	IIIC 80-90	0.982 (0.969, 0.993)	0.855 (0.774, 0.923)	85.2% (73.3%, 96.7%)	84.8% (73.3%, 95.1%)
Morgoth	LRDA	IIIC >90	0.991 (0.976, 1.000)	0.847 (0.559, 1.000)	97.0% (90.0%, 100.0%)	80.5% (68.2%, 90.0%)
Morgoth	Other	IIIC 0-1	0.941 (0.873, 0.995)	0.958 (0.897, 0.996)	51.9% (28.6%, 81.0%)	41.7% (23.8%, 61.9%)
Morgoth	Other	IIIC 1-10	0.967 (0.936, 0.989)	0.934 (0.872, 0.983)	68.3% (51.9%, 90.8%)	57.6% (37.0%, 83.4%)
Morgoth	Other	IIIC 10-20	0.957 (0.930, 0.977)	0.950 (0.920, 0.972)	85.1% (70.0%, 96.7%)	85.3% (73.3%, 95.1%)
Morgoth	Other	IIIC 20-30	0.981 (0.970, 0.990)	0.964 (0.941, 0.983)	91.8% (83.3%, 96.7%)	85.2% (73.3%, 95.1%)
Morgoth	Other	IIIC 30-40	0.988 (0.980, 0.995)	0.982 (0.971, 0.992)	96.1% (93.3%, 100.0%)	92.9% (86.7%, 100.0%)
Morgoth	Other	IIIC 40-50	0.987 (0.978, 0.994)	0.980 (0.961, 0.993)	92.1% (83.3%, 100.0%)	88.2% (78.2%, 96.7%)
Morgoth	Other	IIIC 50-60	0.991 (0.986, 0.996)	0.989 (0.984, 0.994)	96.1% (93.3%, 96.7%)	96.0% (93.3%, 96.7%)
Morgoth	Other	IIIC 60-70	0.989 (0.986, 0.992)	0.984 (0.979, 0.989)	99.4% (96.7%, 100.0%)	98.3% (93.3%, 100.0%)
Morgoth	Other	IIIC 70-80	0.991 (0.988, 0.993)	0.991 (0.987, 0.994)	99.6% (96.7%, 100.0%)	99.6% (96.7%, 100.0%)
Morgoth	Other	IIIC 80-90	0.984 (0.977, 0.991)	0.970 (0.959, 0.980)	96.6% (90.0%, 100.0%)	94.5% (86.7%, 100.0%)
Morgoth	Other	IIIC >90	0.992 (0.983, 0.999)	0.995 (0.989, 0.999)	81.4% (70.0%, 93.3%)	77.1% (66.7%, 86.7%)
Morgoth	Seizure	IIIC 0-1	0.735 (0.459, 0.939)	0.271 (0.064, 0.631)	85.3% (33.3%, 100.0%)	84.3% (33.3%, 100.0%)
Morgoth	Seizure	IIIC 1-10	0.845 (0.771, 0.917)	0.521 (0.379, 0.691)	18.8% (7.4%, 29.6%)	26.2% (9.2%, 44.4%)
Morgoth	Seizure	IIIC 10-20	0.969 (0.951, 0.987)	0.707 (0.490, 0.853)	43.3% (26.7%, 66.7%)	34.3% (16.7%, 50.0%)
Morgoth	Seizure	IIIC 20-30	0.978 (0.967, 0.987)	0.858 (0.786, 0.912)	67.3% (53.3%, 83.3%)	70.7% (54.9%, 81.7%)
Morgoth	Seizure	IIIC 30-40	0.991 (0.984, 0.997)	0.855 (0.740, 0.955)	68.0% (50.0%, 90.0%)	63.1% (46.7%, 83.3%)
Morgoth	Seizure	IIIC 40-50	0.975 (0.945, 0.995)	0.591 (0.328, 0.842)	70.4% (53.3%, 90.0%)	62.9% (40.0%, 83.3%)
Morgoth	Seizure	IIIC 50-60	0.995 (0.991, 0.997)	0.842 (0.719, 0.940)	91.0% (80.0%, 100.0%)	85.6% (73.3%, 93.3%)
Morgoth	Seizure	IIIC 60-70	0.993 (0.989, 0.996)	0.972 (0.954, 0.986)	80.4% (70.0%, 90.0%)	83.0% (73.3%, 90.0%)

Morgoth	Seizure	IIIC 70-80	0.992 (0.986, 0.996)	0.883 (0.817, 0.933)	77.1% (63.3%, 90.0%)	73.1% (56.7%, 83.3%)
Morgoth	Seizure	IIIC 80-90	0.994 (0.990, 0.998)	0.966 (0.934, 0.986)	76.7% (66.7%, 86.7%)	77.2% (66.7%, 86.7%)
Morgoth	Seizure	IIIC >90	0.987 (0.971, 1.000)	0.490 (0.145, 1.000)	100.0% (100.0%, 100.0%)	99.8% (100.0%, 100.0%)
SPaRCNet	GPD	IIIC 1-10	0.789 (0.684, 0.865)	0.668 (0.501, 0.799)	14.7% (5.5%, 25.9%)	14.8% (3.7%, 25.9%)
SPaRCNet	GPD	IIIC 10-20	0.923 (0.854, 0.979)	0.762 (0.597, 0.883)	50.1% (31.6%, 75.1%)	45.8% (26.7%, 66.8%)
SPaRCNet	GPD	IIIC 20-30	0.864 (0.827, 0.898)	0.547 (0.458, 0.639)	19.5% (13.3%, 26.7%)	17.9% (13.3%, 23.3%)
SPaRCNet	GPD	IIIC 30-40	0.943 (0.925, 0.962)	0.878 (0.836, 0.918)	47.6% (28.2%, 66.7%)	49.4% (30.0%, 71.7%)
SPaRCNet	GPD	IIIC 40-50	0.790 (0.683, 0.877)	0.589 (0.422, 0.722)	40.3% (23.3%, 60.0%)	39.2% (20.0%, 60.0%)
SPaRCNet	GPD	IIIC 50-60	0.907 (0.881, 0.926)	0.778 (0.724, 0.826)	35.2% (20.0%, 51.7%)	33.9% (16.7%, 48.4%)
SPaRCNet	GPD	IIIC 60-70	0.914 (0.880, 0.943)	0.665 (0.578, 0.746)	46.6% (21.6%, 73.3%)	29.3% (10.0%, 53.3%)
SPaRCNet	GPD	IIIC 70-80	0.944 (0.923, 0.962)	0.707 (0.626, 0.783)	33.2% (16.7%, 53.3%)	32.7% (18.2%, 50.0%)
SPaRCNet	GPD	IIIC 80-90	0.979 (0.967, 0.989)	0.883 (0.838, 0.923)	63.9% (46.7%, 80.0%)	53.6% (36.7%, 73.3%)
SPaRCNet	GPD	IIIC >90	0.929 (0.854, 0.989)	0.555 (0.284, 0.813)	61.6% (46.7%, 78.4%)	54.8% (43.3%, 66.7%)
SPaRCNet	GRDA	IIIC 1-10	0.901 (0.789, 0.978)	0.435 (0.112, 0.705)	57.1% (40.7%, 81.5%)	52.1% (37.0%, 70.4%)
SPaRCNet	GRDA	IIIC 10-20	0.919 (0.861, 0.964)	0.689 (0.532, 0.824)	40.7% (23.3%, 56.7%)	42.4% (26.7%, 60.0%)
SPaRCNet	GRDA	IIIC 20-30	0.851 (0.807, 0.886)	0.639 (0.536, 0.712)	24.1% (13.3%, 36.7%)	25.3% (13.3%, 36.7%)
SPaRCNet	GRDA	IIIC 30-40	0.892 (0.861, 0.915)	0.706 (0.633, 0.772)	17.3% (6.7%, 30.0%)	15.4% (4.9%, 26.7%)
SPaRCNet	GRDA	IIIC 40-50	0.818 (0.756, 0.867)	0.512 (0.406, 0.624)	25.7% (13.3%, 40.0%)	23.8% (11.6%, 40.0%)
SPaRCNet	GRDA	IIIC 50-60	0.909 (0.868, 0.944)	0.415 (0.285, 0.552)	32.2% (16.7%, 56.8%)	36.2% (18.2%, 60.0%)
SPaRCNet	GRDA	IIIC 60-70	0.912 (0.877, 0.939)	0.470 (0.387, 0.557)	29.0% (13.3%, 46.7%)	20.1% (3.3%, 38.4%)
SPaRCNet	GRDA	IIIC 70-80	0.891 (0.862, 0.911)	0.458 (0.381, 0.531)	15.0% (3.3%, 28.4%)	19.2% (6.7%, 35.1%)
SPaRCNet	GRDA	IIIC 80-90	0.964 (0.949, 0.976)	0.513 (0.338, 0.655)	40.3% (26.7%, 60.0%)	33.7% (18.2%, 55.1%)
SPaRCNet	GRDA	IIIC >90	0.907 (0.859, 0.953)	0.559 (0.411, 0.745)	73.1% (60.0%, 83.3%)	66.8% (50.0%, 81.7%)
SPaRCNet	LPD	IIIC 0-1	0.868 (0.714, 0.983)	0.723 (0.301, 0.967)	43.7% (14.3%, 83.5%)	39.3% (14.3%, 76.2%)
SPaRCNet	LPD	IIIC 1-10	0.847 (0.719, 0.953)	0.569 (0.328, 0.788)	25.1% (14.8%, 40.7%)	25.8% (14.8%, 42.7%)
SPaRCNet	LPD	IIIC 10-20	0.842 (0.790, 0.889)	0.737 (0.653, 0.809)	19.8% (6.7%, 33.3%)	18.8% (6.7%, 36.7%)
SPaRCNet	LPD	IIIC 20-30	0.822 (0.784, 0.861)	0.506 (0.434, 0.579)	15.0% (10.0%, 23.3%)	17.2% (10.0%, 26.7%)
SPaRCNet	LPD	IIIC 30-40	0.699 (0.495, 0.902)	0.118 (0.031, 0.279)	49.0% (41.6%, 60.0%)	48.7% (40.0%, 56.7%)
SPaRCNet	LPD	IIIC 40-50	0.842 (0.812, 0.881)	0.639 (0.557, 0.721)	7.7% (0.0%, 16.7%)	11.8% (3.3%, 26.7%)
SPaRCNet	LPD	IIIC 50-60	0.894 (0.879, 0.914)	0.740 (0.701, 0.785)	8.6% (1.6%, 18.4%)	11.7% (3.3%, 20.0%)
SPaRCNet	LPD	IIIC 60-70	0.733 (0.712, 0.749)	0.553 (0.514, 0.582)	0.8% (0.0%, 3.3%)	3.3% (3.3%, 3.3%)
SPaRCNet	LPD	IIIC 70-80	0.901 (0.880, 0.921)	0.760 (0.717, 0.822)	17.4% (4.9%, 30.0%)	14.6% (3.3%, 28.4%)
SPaRCNet	LPD	IIIC 80-90	0.901 (0.886, 0.915)	0.887 (0.866, 0.904)	31.1% (16.7%, 43.3%)	49.0% (34.9%, 60.0%)
SPaRCNet	LPD	IIIC >90	0.962 (0.937, 0.985)	0.915 (0.863, 0.958)	31.8% (20.0%, 50.0%)	34.3% (20.0%, 51.7%)
SPaRCNet	LRDA	IIIC 1-10	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	97.8% (69.4%, 100.0%)
SPaRCNet	LRDA	IIIC 10-20	0.981 (0.957, 0.992)	0.262 (0.106, 0.536)	99.3% (88.3%, 100.0%)	99.1% (81.7%, 100.0%)
SPaRCNet	LRDA	IIIC 20-30	0.839 (0.795, 0.892)	0.371 (0.235, 0.491)	50.8% (36.7%, 66.7%)	57.5% (41.6%, 75.1%)
SPaRCNet	LRDA	IIIC 30-40	0.989 (0.975, 1.000)	0.598 (0.102, 1.000)	72.2% (50.0%, 100.0%)	50.6% (40.0%, 100.0%)
SPaRCNet	LRDA	IIIC 40-50	0.761 (0.686, 0.821)	0.278 (0.160, 0.412)	16.0% (6.7%, 25.1%)	17.9% (10.0%, 30.0%)
SPaRCNet	LRDA	IIIC 50-60	0.944 (0.921, 0.965)	0.672 (0.590, 0.760)	61.0% (40.0%, 80.0%)	60.1% (43.3%, 80.0%)
SPaRCNet	LRDA	IIIC 60-70	0.788 (0.735, 0.844)	0.125 (0.084, 0.196)	20.5% (10.0%, 38.4%)	17.8% (6.7%, 31.7%)
SPaRCNet	LRDA	IIIC 70-80	0.869 (0.837, 0.899)	0.446 (0.351, 0.532)	15.3% (0.0%, 35.1%)	17.7% (4.9%, 36.7%)
SPaRCNet	LRDA	IIIC 80-90	0.851 (0.805, 0.888)	0.224 (0.137, 0.313)	46.3% (31.6%, 61.7%)	45.1% (28.2%, 61.7%)

SPaRCNet	LRDA	IIIC >90	0.871 (0.784, 0.945)	0.201 (0.066, 0.423)	64.8% (53.3%, 83.3%)	59.6% (50.0%, 71.7%)
SPaRCNet	Other	IIIC 0-1	0.827 (0.687, 0.927)	0.874 (0.732, 0.949)	22.6% (4.8%, 50.1%)	21.9% (4.8%, 42.9%)
SPaRCNet	Other	IIIC 1-10	0.825 (0.767, 0.893)	0.652 (0.492, 0.815)	21.7% (11.1%, 39.0%)	15.9% (11.1%, 31.6%)
SPaRCNet	Other	IIIC 10-20	0.769 (0.710, 0.828)	0.664 (0.562, 0.747)	18.7% (3.3%, 36.7%)	25.1% (8.2%, 41.7%)
SPaRCNet	Other	IIIC 20-30	0.814 (0.777, 0.847)	0.697 (0.617, 0.762)	6.1% (0.0%, 13.3%)	5.9% (0.0%, 13.3%)
SPaRCNet	Other	IIIC 30-40	0.832 (0.797, 0.871)	0.740 (0.676, 0.812)	13.8% (3.3%, 26.7%)	16.1% (3.3%, 31.7%)
SPaRCNet	Other	IIIC 40-50	0.718 (0.684, 0.751)	0.624 (0.576, 0.689)	3.6% (0.0%, 10.0%)	7.0% (0.0%, 16.7%)
SPaRCNet	Other	IIIC 50-60	0.816 (0.792, 0.840)	0.791 (0.753, 0.823)	5.5% (0.0%, 13.3%)	5.2% (0.0%, 13.3%)
SPaRCNet	Other	IIIC 60-70	0.701 (0.679, 0.720)	0.567 (0.537, 0.602)	6.2% (3.3%, 10.0%)	2.8% (0.0%, 6.7%)
SPaRCNet	Other	IIIC 70-80	0.815 (0.797, 0.841)	0.797 (0.770, 0.832)	4.5% (0.0%, 11.7%)	12.6% (4.9%, 21.7%)
SPaRCNet	Other	IIIC 80-90	0.831 (0.815, 0.851)	0.613 (0.570, 0.658)	15.9% (8.2%, 26.7%)	5.3% (0.0%, 13.3%)
SPaRCNet	Other	IIIC >90	0.777 (0.705, 0.833)	0.815 (0.753, 0.863)	11.4% (3.3%, 20.0%)	8.8% (1.6%, 21.7%)
SPaRCNet	Seizure	IIIC 0-1	0.751 (0.516, 0.939)	0.282 (0.067, 0.600)	85.1% (33.3%, 100.0%)	84.3% (33.3%, 100.0%)
SPaRCNet	Seizure	IIIC 1-10	0.610 (0.506, 0.739)	0.341 (0.187, 0.508)	13.6% (3.7%, 22.2%)	17.0% (5.5%, 29.6%)
SPaRCNet	Seizure	IIIC 10-20	0.877 (0.821, 0.927)	0.442 (0.222, 0.660)	22.5% (13.3%, 33.3%)	22.6% (13.3%, 33.3%)
SPaRCNet	Seizure	IIIC 20-30	0.854 (0.818, 0.888)	0.519 (0.398, 0.615)	21.4% (10.0%, 35.1%)	25.9% (11.6%, 38.4%)
SPaRCNet	Seizure	IIIC 30-40	0.913 (0.876, 0.946)	0.441 (0.290, 0.629)	24.8% (10.0%, 46.7%)	25.9% (8.2%, 48.4%)
SPaRCNet	Seizure	IIIC 40-50	0.672 (0.513, 0.836)	0.118 (0.031, 0.336)	41.0% (30.0%, 58.4%)	39.5% (28.2%, 58.4%)
SPaRCNet	Seizure	IIIC 50-60	0.700 (0.632, 0.757)	0.072 (0.051, 0.094)	13.0% (6.7%, 20.0%)	13.0% (6.7%, 16.7%)
SPaRCNet	Seizure	IIIC 60-70	0.854 (0.838, 0.867)	0.417 (0.381, 0.455)	21.4% (16.7%, 28.4%)	49.8% (40.0%, 60.0%)
SPaRCNet	Seizure	IIIC 70-80	0.939 (0.919, 0.958)	0.434 (0.318, 0.565)	13.5% (3.3%, 26.7%)	12.7% (1.6%, 25.1%)
SPaRCNet	Seizure	IIIC 80-90	0.954 (0.942, 0.966)	0.689 (0.615, 0.761)	25.0% (13.3%, 41.7%)	49.4% (33.3%, 70.0%)
SPaRCNet	Seizure	IIIC >90	0.918 (0.808, 1.000)	0.529 (0.026, 1.000)	99.9% (100.0%, 100.0%)	99.8% (100.0%, 100.0%)
Kaggle Winner	GPD	IIIC 1-10	0.956 (0.922, 0.980)	0.890 (0.816, 0.956)	47.6% (22.2%, 74.1%)	45.8% (22.2%, 72.3%)
Kaggle Winner	GPD	IIIC 10-20	0.991 (0.977, 0.999)	0.936 (0.861, 0.990)	76.2% (54.9%, 95.1%)	71.8% (50.0%, 93.3%)
Kaggle Winner	GPD	IIIC 20-30	0.982 (0.972, 0.990)	0.907 (0.859, 0.945)	79.8% (64.9%, 93.3%)	76.0% (54.9%, 90.0%)
Kaggle Winner	GPD	IIIC 30-40	0.965 (0.945, 0.979)	0.929 (0.898, 0.955)	73.9% (51.6%, 91.7%)	73.7% (54.9%, 91.7%)
Kaggle Winner	GPD	IIIC 40-50	0.951 (0.916, 0.980)	0.808 (0.702, 0.910)	66.1% (46.7%, 86.7%)	63.2% (44.9%, 83.3%)
Kaggle Winner	GPD	IIIC 50-60	0.973 (0.962, 0.981)	0.892 (0.855, 0.920)	61.8% (41.6%, 76.7%)	58.3% (36.7%, 76.7%)
Kaggle Winner	GPD	IIIC 60-70	0.973 (0.953, 0.988)	0.817 (0.747, 0.871)	83.6% (64.9%, 96.7%)	67.7% (40.0%, 86.7%)
Kaggle Winner	GPD	IIIC 70-80	0.979 (0.973, 0.986)	0.820 (0.768, 0.871)	65.9% (43.3%, 90.0%)	58.2% (40.0%, 78.4%)
Kaggle Winner	GPD	IIIC 80-90	0.992 (0.988, 0.995)	0.923 (0.876, 0.960)	82.4% (66.7%, 93.3%)	71.6% (49.8%, 86.7%)
Kaggle Winner	GPD	IIIC >90	0.961 (0.903, 0.994)	0.649 (0.296, 0.897)	65.6% (48.2%, 85.1%)	57.9% (41.6%, 71.7%)
Kaggle Winner	GRDA	IIIC 1-10	0.954 (0.891, 0.997)	0.670 (0.277, 0.958)	66.2% (49.9%, 94.7%)	59.4% (44.4%, 76.0%)
Kaggle Winner	GRDA	IIIC 10-20	0.971 (0.954, 0.988)	0.830 (0.732, 0.931)	65.1% (50.0%, 80.0%)	55.3% (36.7%, 73.3%)
Kaggle Winner	GRDA	IIIC 20-30	0.937 (0.915, 0.954)	0.774 (0.693, 0.835)	49.1% (30.0%, 68.4%)	51.1% (31.6%, 71.7%)
Kaggle Winner	GRDA	IIIC 30-40	0.958 (0.944, 0.971)	0.872 (0.815, 0.915)	53.8% (34.9%, 73.3%)	47.2% (26.7%, 65.1%)
Kaggle Winner	GRDA	IIIC 40-50	0.956 (0.938, 0.969)	0.770 (0.681, 0.848)	55.3% (40.0%, 71.7%)	51.8% (33.3%, 75.1%)
Kaggle Winner	GRDA	IIIC 50-60	0.963 (0.937, 0.983)	0.633 (0.496, 0.767)	69.4% (46.5%, 86.7%)	68.8% (46.7%, 86.7%)
Kaggle Winner	GRDA	IIIC 60-70	0.966 (0.952, 0.978)	0.684 (0.603, 0.763)	62.4% (50.0%, 76.7%)	56.1% (40.0%, 71.7%)
Kaggle Winner	GRDA	IIIC 70-80	0.950 (0.934, 0.962)	0.662 (0.557, 0.745)	50.5% (30.0%, 66.7%)	54.7% (29.8%, 73.3%)
Kaggle Winner	GRDA	IIIC 80-90	0.981 (0.967, 0.991)	0.693 (0.559, 0.805)	64.7% (43.2%, 83.3%)	59.7% (34.9%, 80.0%)
Kaggle Winner	GRDA	IIIC >90	0.921 (0.887, 0.957)	0.521 (0.362, 0.742)	75.3% (60.0%, 90.0%)	67.4% (50.0%, 83.3%)

Kaggle Winner	LPD	IIIC 0-1	0.922 (0.801, 1.000)	0.802 (0.477, 1.000)	59.9% (21.3%, 100.0%)	50.6% (19.0%, 88.2%)
Kaggle Winner	LPD	IIIC 1-10	0.986 (0.966, 0.998)	0.918 (0.793, 0.982)	66.9% (40.7%, 92.6%)	55.9% (33.3%, 76.0%)
Kaggle Winner	LPD	IIIC 10-20	0.971 (0.951, 0.985)	0.939 (0.898, 0.967)	71.8% (51.6%, 88.4%)	69.4% (50.0%, 85.1%)
Kaggle Winner	LPD	IIIC 20-30	0.960 (0.945, 0.972)	0.832 (0.775, 0.879)	52.0% (34.9%, 68.4%)	60.4% (41.6%, 75.1%)
Kaggle Winner	LPD	IIIC 30-40	0.976 (0.959, 0.993)	0.534 (0.220, 0.763)	62.0% (46.7%, 80.0%)	61.1% (46.7%, 78.4%)
Kaggle Winner	LPD	IIIC 40-50	0.949 (0.930, 0.964)	0.867 (0.807, 0.903)	38.9% (20.0%, 58.4%)	45.8% (21.6%, 66.7%)
Kaggle Winner	LPD	IIIC 50-60	0.970 (0.959, 0.977)	0.920 (0.896, 0.940)	65.5% (48.2%, 80.0%)	69.3% (51.6%, 85.1%)
Kaggle Winner	LPD	IIIC 60-70	0.919 (0.911, 0.929)	0.821 (0.798, 0.844)	21.3% (13.3%, 33.3%)	27.2% (13.3%, 40.0%)
Kaggle Winner	LPD	IIIC 70-80	0.982 (0.975, 0.989)	0.951 (0.938, 0.968)	91.4% (83.3%, 96.7%)	88.0% (76.7%, 96.7%)
Kaggle Winner	LPD	IIIC 80-90	0.883 (0.867, 0.897)	0.838 (0.817, 0.860)	12.1% (4.9%, 21.7%)	29.1% (16.7%, 46.7%)
Kaggle Winner	LPD	IIIC >90	0.987 (0.976, 0.997)	0.972 (0.949, 0.992)	54.2% (36.7%, 73.5%)	52.0% (34.9%, 66.7%)
Kaggle Winner	LRDA	IIIC 1-10	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	97.8% (69.4%, 100.0%)
Kaggle Winner	LRDA	IIIC 10-20	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	99.1% (81.7%, 100.0%)
Kaggle Winner	LRDA	IIIC 20-30	0.949 (0.917, 0.971)	0.659 (0.538, 0.766)	74.2% (56.7%, 88.4%)	82.3% (66.7%, 93.3%)
Kaggle Winner	LRDA	IIIC 30-40	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	57.1% (46.7%, 100.0%)
Kaggle Winner	LRDA	IIIC 40-50	0.909 (0.871, 0.938)	0.497 (0.354, 0.646)	37.6% (20.0%, 60.0%)	37.7% (20.0%, 56.7%)
Kaggle Winner	LRDA	IIIC 50-60	0.972 (0.959, 0.983)	0.790 (0.707, 0.858)	81.5% (64.9%, 95.1%)	79.2% (60.0%, 93.3%)
Kaggle Winner	LRDA	IIIC 60-70	0.910 (0.868, 0.937)	0.310 (0.212, 0.396)	50.3% (31.6%, 68.4%)	44.2% (26.7%, 66.7%)
Kaggle Winner	LRDA	IIIC 70-80	0.952 (0.934, 0.966)	0.678 (0.597, 0.754)	66.8% (46.5%, 83.3%)	67.2% (41.6%, 85.1%)
Kaggle Winner	LRDA	IIIC 80-90	0.871 (0.830, 0.917)	0.319 (0.197, 0.445)	59.2% (46.7%, 71.7%)	58.9% (43.3%, 73.3%)
Kaggle Winner	LRDA	IIIC >90	0.935 (0.894, 0.985)	0.387 (0.115, 0.749)	77.5% (61.6%, 93.3%)	64.8% (53.3%, 80.0%)
Kaggle Winner	Other	IIIC 0-1	0.965 (0.895, 1.000)	0.965 (0.888, 1.000)	66.9% (33.3%, 100.0%)	40.0% (16.5%, 66.7%)
Kaggle Winner	Other	IIIC 1-10	0.936 (0.898, 0.979)	0.872 (0.768, 0.966)	54.2% (35.1%, 81.5%)	40.4% (22.2%, 70.4%)
Kaggle Winner	Other	IIIC 10-20	0.909 (0.868, 0.939)	0.894 (0.845, 0.929)	66.1% (46.7%, 83.3%)	69.5% (50.0%, 86.7%)
Kaggle Winner	Other	IIIC 20-30	0.955 (0.941, 0.967)	0.917 (0.888, 0.938)	71.4% (56.7%, 88.4%)	50.1% (28.2%, 68.4%)
Kaggle Winner	Other	IIIC 30-40	0.952 (0.927, 0.970)	0.922 (0.879, 0.958)	81.8% (61.6%, 93.3%)	77.7% (60.0%, 90.0%)
Kaggle Winner	Other	IIIC 40-50	0.907 (0.888, 0.927)	0.899 (0.876, 0.926)	55.0% (40.0%, 71.7%)	50.2% (36.7%, 66.7%)
Kaggle Winner	Other	IIIC 50-60	0.944 (0.931, 0.955)	0.927 (0.901, 0.948)	78.1% (61.6%, 93.3%)	71.5% (53.2%, 88.4%)
Kaggle Winner	Other	IIIC 60-70	0.855 (0.836, 0.868)	0.831 (0.808, 0.849)	27.9% (18.2%, 36.7%)	15.7% (10.0%, 23.3%)
Kaggle Winner	Other	IIIC 70-80	0.946 (0.937, 0.955)	0.945 (0.933, 0.955)	74.7% (58.2%, 86.7%)	76.1% (58.2%, 86.7%)
Kaggle Winner	Other	IIIC 80-90	0.887 (0.871, 0.903)	0.817 (0.784, 0.846)	42.2% (33.3%, 50.0%)	28.8% (16.7%, 41.7%)
Kaggle Winner	Other	IIIC >90	0.898 (0.859, 0.931)	0.922 (0.888, 0.954)	34.4% (20.0%, 51.7%)	30.9% (13.3%, 53.3%)
Kaggle Winner	Seizure	IIIC 0-1	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	87.4% (42.9%, 100.0%)
Kaggle Winner	Seizure	IIIC 1-10	0.845 (0.741, 0.937)	0.670 (0.488, 0.838)	27.2% (11.1%, 48.1%)	32.8% (12.9%, 51.9%)
Kaggle Winner	Seizure	IIIC 10-20	0.963 (0.926, 0.994)	0.855 (0.704, 0.955)	54.4% (33.3%, 80.0%)	51.5% (31.6%, 71.7%)
Kaggle Winner	Seizure	IIIC 20-30	0.913 (0.884, 0.935)	0.687 (0.598, 0.758)	36.6% (20.0%, 55.1%)	39.9% (20.0%, 56.7%)
Kaggle Winner	Seizure	IIIC 30-40	0.970 (0.944, 0.988)	0.769 (0.615, 0.877)	57.8% (33.3%, 78.4%)	57.3% (36.7%, 76.7%)
Kaggle Winner	Seizure	IIIC 40-50	0.949 (0.911, 0.982)	0.506 (0.287, 0.767)	59.4% (36.7%, 80.0%)	53.9% (33.3%, 75.1%)
Kaggle Winner	Seizure	IIIC 50-60	0.964 (0.944, 0.980)	0.693 (0.585, 0.793)	51.3% (33.3%, 68.4%)	56.2% (41.6%, 70.0%)
Kaggle Winner	Seizure	IIIC 60-70	0.956 (0.949, 0.964)	0.802 (0.771, 0.841)	49.6% (40.0%, 60.0%)	75.3% (66.7%, 83.3%)
Kaggle Winner	Seizure	IIIC 70-80	0.978 (0.958, 0.992)	0.820 (0.732, 0.896)	66.5% (50.0%, 80.0%)	64.4% (44.9%, 80.0%)
Kaggle Winner	Seizure	IIIC 80-90	0.982 (0.974, 0.990)	0.917 (0.888, 0.940)	64.7% (53.3%, 76.7%)	68.0% (60.0%, 78.4%)
Kaggle Winner	Seizure	IIIC >90	0.929 (0.853, 0.991)	0.191 (0.030, 0.592)	99.9% (100.0%, 100.0%)	99.8% (100.0%, 100.0%)

1168
1169

(B) Internal SN2-Test dataset

Model	Age Group	ROC AUC	ROC EUC	PR AUC	PR EUC
Morgoth	0-1	0.997 (0.981, 1.000)	98.0% (85.7%, 100.0%)	0.998 (0.991, 1.000)	86.4% (76.2%, 90.5%)
SpikeNet2	0-1	0.990 (0.958, 1.000)	95.1% (81.0%, 100.0%)	0.995 (0.978, 1.000)	84.0% (71.4%, 90.5%)
Morgoth	1-10	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	83.8% (70.8%, 91.7%)
SpikeNet2	1-10	0.984 (0.961, 0.997)	56.6% (33.3%, 79.2%)	0.995 (0.987, 0.999)	59.4% (37.5%, 83.3%)
Morgoth	10-20	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	64.5% (54.2%, 70.8%)
SpikeNet2	10-20	0.972 (0.938, 0.994)	40.8% (20.8%, 66.7%)	0.993 (0.984, 0.999)	40.1% (20.8%, 66.7%)
Morgoth	20-30	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	51.0% (41.7%, 62.5%)
SpikeNet2	20-30	0.955 (0.883, 1.000)	25.7% (4.2%, 100.0%)	0.994 (0.983, 1.000)	22.4% (8.3%, 54.2%)
Morgoth	30-40	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	50.7% (38.1%, 66.7%)
SpikeNet2	30-40	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	50.7% (38.1%, 66.7%)
Morgoth	40-50	nan (nan, nan)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	85.0% (79.2%, 87.5%)
SpikeNet2	40-50	nan (nan, nan)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	85.0% (79.2%, 87.5%)
Morgoth	50-60	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	87.3% (71.4%, 100.0%)
SpikeNet2	50-60	0.979 (0.932, 1.000)	82.4% (61.9%, 100.0%)	0.990 (0.969, 1.000)	75.2% (52.4%, 95.2%)
Morgoth	60-70	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	47.8% (37.5%, 54.3%)
SpikeNet2	60-70	0.981 (0.950, 0.998)	43.9% (20.8%, 79.2%)	0.988 (0.971, 0.999)	34.7% (16.7%, 50.0%)
Morgoth	70-80	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	68.0% (58.3%, 79.2%)
SpikeNet2	70-80	0.955 (0.895, 1.000)	73.2% (50.0%, 100.0%)	0.943 (0.862, 1.000)	51.1% (33.3%, 70.8%)
Morgoth	>80	0.982 (0.918, 1.000)	85.1% (54.2%, 100.0%)	0.975 (0.889, 1.000)	61.5% (45.7%, 70.8%)
SpikeNet2	>80	0.980 (0.911, 1.000)	84.7% (54.2%, 100.0%)	0.974 (0.881, 1.000)	61.5% (45.8%, 70.8%)

1170
1171

(C) Internal MoE dataset

Task	Age Group	ROC AUC	PR AUC	ROC EUC	PR EUC
Burst Suppression	0-20	0.907 (0.860, 0.941)	0.782 (0.667, 0.862)	77.0% (47.6%, 100.0%)	78.5% (45.5%, 100.0%)
Burst Suppression	20-30	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	59.0% (45.5%, 72.7%)	74.1% (33.8%, 86.4%)
Burst Suppression	30-50	0.999(0.997, 1.000)	0.973(0.922, 1.000)	67.1%(50.0%, 81.8%)	77.5%(51.9%, 90.9%)
Burst Suppression	50-60	1.000(1.000, 1.000)	1.000(1.000, 1.000)	73.3%(54.5%, 88.7%)	71.7%(52.2%, 77.3%)
Burst Suppression	60-70	0.999(0.997, 1.000)	0.989(0.952, 1.000)	82.9%(68.2%, 90.9%)	84.1%(63.6%, 95.5%)
Burst Suppression	>70	0.997(0.992, 1.000)	0.895(0.725, 1.000)	76.2%(63.6%, 86.4%)	81.4%(61.2%, 95.5%)
Focal Slowing	0-20	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	89.1% (77.3%, 100.0%)	82.3% (72.7%, 100.0%)
Focal Slowing	20-30	0.930 (0.876, 0.975)	0.791 (0.633, 0.908)	68.2% (43.1%, 86.4%)	76.6% (54.5%, 95.5%)
Focal Slowing	30-50	0.925(0.880, 0.959)	0.839(0.762, 0.902)	82.7%(65.8%, 95.5%)	86.6%(72.7%, 95.5%)
Focal Slowing	50-60	0.906(0.853, 0.949)	0.851(0.786, 0.917)	84.5%(72.7%, 95.5%)	89.6%(81.8%, 100.0%)
Focal Slowing	60-70	0.946(0.912, 0.972)	0.902(0.848, 0.949)	87.2%(72.7%, 100.0%)	95.0%(86.4%, 100.0%)
Focal Slowing	>70	0.946(0.915, 0.967)	0.907(0.865, 0.945)	87.0%(81.8%, 90.9%)	96.1%(90.9%, 100.0%)
Focal Spikes	0-20	0.844 (0.776, 0.893)	0.842 (0.767, 0.917)	86.4% (72.7%, 100.0%)	84.6% (68.2%, 100.0%)
Focal Spikes	20-30	0.951 (0.897, 0.993)	0.797 (0.587, 0.970)	77.4% (59.1%, 95.5%)	81.5% (59.1%, 95.5%)
Focal Spikes	30-50	0.966(0.918, 0.993)	0.915(0.826, 0.968)	84.7%(68.2%, 95.5%)	91.3%(79.4%, 100.0%)
Focal Spikes	50-60	0.949(0.914, 0.977)	0.783(0.657, 0.891)	58.1%(31.8%, 77.3%)	66.8%(47.6%, 86.4%)

Focal Spikes	60-70	0.966(0.933, 0.996)	0.885(0.792, 0.970)	69.4%(43.1%, 86.4%)	82.8%(54.5%, 95.5%)
Focal Spikes	>70	0.972(0.943, 0.987)	0.888(0.798, 0.945)	77.6%(59.1%, 90.9%)	89.0%(72.7%, 100.0%)
Generalized Slowing	0-20	0.686 (0.598, 0.767)	0.172 (0.080, 0.316)	25.2% (13.6%, 47.8%)	25.7% (13.6%, 47.8%)
Generalized Slowing	20-30	0.859 (0.815, 0.910)	0.748 (0.640, 0.824)	74.8% (54.5%, 90.9%)	75.2% (54.5%, 95.5%)
Generalized Slowing	30-50	0.836(0.804, 0.872)	0.755(0.687, 0.811)	79.9%(63.6%, 93.3%)	79.5%(63.6%, 95.5%)
Generalized Slowing	50-60	0.883(0.852, 0.917)	0.860(0.818, 0.906)	95.5%(84.0%, 100.0%)	96.5%(86.4%, 100.0%)
Generalized Slowing	60-70	0.885(0.842, 0.916)	0.886(0.844, 0.919)	95.1%(86.4%, 100.0%)	96.1%(88.5%, 100.0%)
Generalized Slowing	>70	0.881(0.847, 0.907)	0.900(0.875, 0.924)	95.2%(86.4%, 100.0%)	97.3%(90.9%, 100.0%)
Generalized Spikes	0-20	0.905 (0.856, 0.956)	0.754 (0.617, 0.880)	73.2% (40.9%, 95.5%)	76.9% (43.1%, 100.0%)
Generalized Spikes	20-30	0.964 (0.907, 0.997)	0.751 (0.426, 0.948)	41.9% (18.2%, 66.0%)	48.3% (22.7%, 77.3%)
Generalized Spikes	30-50	0.935(0.864, 0.994)	0.617(0.385, 0.853)	53.4%(22.7%, 79.7%)	61.9%(27.3%, 88.7%)
Generalized Spikes	50-60	0.973(0.951, 0.995)	0.789(0.599, 0.935)	78.6%(52.2%, 93.3%)	88.4%(68.2%, 100.0%)
Generalized Spikes	60-70	0.992(0.983, 0.999)	0.895(0.776, 0.986)	51.6%(27.3%, 72.7%)	73.9%(36.4%, 100.0%)
Generalized Spikes	>70	0.966(0.925, 0.993)	0.784(0.594, 0.931)	63.0%(29.4%, 81.8%)	81.4%(50.0%, 97.8%)
IIIC - GPD	0-20	0.991 (0.972, 1.000)	0.889 (0.680, 1.000)	74.0% (47.6%, 95.5%)	84.6% (54.5%, 100.0%)
IIIC - GPD	20-30	0.962 (0.905, 1.000)	0.555 (0.059, 1.000)	59.7% (36.4%, 100.0%)	59.4% (36.4%, 100.0%)
IIIC - GPD	30-50	0.959(0.915, 0.997)	0.823(0.512, 0.992)	40.5%(20.3%, 63.9%)	62.4%(34.0%, 79.7%)
IIIC - GPD	50-60	0.945(0.800, 1.000)	0.802(0.484, 1.000)	62.3%(36.4%, 88.7%)	73.5%(50.0%, 90.9%)
IIIC - GPD	60-70	0.990(0.978, 0.999)	0.872(0.676, 0.985)	58.5%(36.4%, 77.3%)	84.3%(54.5%, 100.0%)
IIIC - GPD	>70	0.965(0.910, 0.998)	0.799(0.532, 0.971)	61.9%(38.5%, 86.4%)	81.6%(63.6%, 97.8%)
IIIC - GRDA	0-20	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	73.2% (54.5%, 100.0%)	69.7% (54.5%, 100.0%)
IIIC - GRDA	20-30	0.964 (0.942, 0.990)	0.195 (0.083, 0.558)	100.0% (100.0%, 100.0%)	100.0% (100.0%, 100.0%)
IIIC - GRDA	30-50	1.000(1.000, 1.000)	1.000(1.000, 1.000)	77.9%(54.5%, 100.0%)	69.6%(45.5%, 100.0%)
IIIC - GRDA	50-60	0.989(0.960, 1.000)	0.868(0.365, 1.000)	74.6%(63.6%, 93.5%)	80.0%(59.1%, 95.7%)
IIIC - GRDA	60-70	1.000(1.000, 1.000)	1.000(1.000, 1.000)	91.4%(81.8%, 100.0%)	83.6%(72.7%, 100.0%)
IIIC - GRDA	>70	0.978(0.937, 1.000)	0.692(0.088, 1.000)	59.4%(36.4%, 93.5%)	63.9%(38.5%, 93.5%)
IIIC - LPD	0-20	0.988 (0.964, 1.000)	0.911 (0.797, 1.000)	68.9% (50.0%, 84.2%)	79.1% (68.2%, 90.9%)
IIIC - LPD	20-30	0.997 (0.990, 1.000)	0.941 (0.716, 1.000)	55.3% (34.0%, 75.1%)	72.6% (38.5%, 86.4%)
IIIC - LPD	30-50	0.906(0.770, 1.000)	0.849(0.628, 1.000)	54.4%(36.4%, 72.7%)	85.6%(61.2%, 95.5%)
IIIC - LPD	50-60	0.982(0.945, 1.000)	0.912(0.727, 1.000)	73.0%(50.0%, 86.4%)	91.1%(68.2%, 100.0%)
IIIC - LPD	60-70	0.962(0.911, 0.996)	0.836(0.687, 0.948)	72.5%(49.8%, 90.9%)	87.0%(63.6%, 100.0%)
IIIC - LPD	>70	0.968(0.938, 0.991)	0.845(0.724, 0.939)	81.4%(63.6%, 90.9%)	88.1%(72.7%, 100.0%)
IIIC - LRDA	0-20	0.987 (0.972, 0.998)	0.444 (0.145, 0.917)	82.1% (63.6%, 100.0%)	66.6% (52.2%, 100.0%)
IIIC - LRDA	20-30	0.972 (0.949, 0.993)	0.240 (0.100, 0.604)	100.0% (100.0%, 100.0%)	100.0% (100.0%, 100.0%)
IIIC - LRDA	30-50	0.982(0.949, 1.000)	0.792(0.369, 1.000)	47.7%(31.8%, 70.6%)	80.0%(54.5%, 90.9%)
IIIC - LRDA	50-60	0.997(0.987, 1.000)	0.974(0.895, 1.000)	75.7%(59.1%, 86.4%)	88.7%(68.2%, 95.5%)
IIIC - LRDA	60-70	0.987(0.966, 1.000)	0.604(0.113, 1.000)	83.7%(70.3%, 100.0%)	77.1%(68.2%, 100.0%)
IIIC - LRDA	>70	0.933(0.810, 1.000)	0.812(0.579, 1.000)	87.2%(72.7%, 100.0%)	95.9%(81.8%, 100.0%)
IIIC - Seizure	0-20	0.988(0.964, 1.000)	0.911(0.797, 1.000)	68.9%(50.0%, 84.2%)	79.1%(68.2%, 90.9%)
IIIC - Seizure	20-30	0.942(0.895, 0.994)	0.479(0.057, 0.928)	44.9%(22.7%, 79.7%)	56.3%(24.9%, 86.4%)
IIIC - Seizure	30-50	1.000(1.000, 1.000)	1.000(1.000, 1.000)	58.8%(22.7%, 90.9%)	79.8%(59.1%, 86.4%)

IIIC - Seizure	50-60	0.896(0.698, 1.000)	0.787(0.454, 1.000)	65.3%(47.6%, 100.0%)	82.0%(50.0%, 100.0%)
IIIC - Seizure	60-70	0.973(0.942, 1.000)	0.824(0.624, 1.000)	64.7%(50.0%, 81.8%)	95.5%(77.3%, 100.0%)
IIIC - Seizure	>70	0.921(0.829, 0.981)	0.358(0.035, 0.719)	49.6%(29.4%, 77.3%)	58.0%(40.9%, 79.7%)
Sleep - Awake	0-20	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)	56.7% (34.0%, 100.0%)	46.4% (31.8%, 100.0%)
Sleep - Awake	20-30	0.993 (0.981, 1.000)	0.997 (0.990, 1.000)	65.4% (52.2%, 77.3%)	85.1% (77.3%, 90.9%)
Sleep - Awake	30-50	0.995(0.985, 1.000)	0.998(0.995, 1.000)	81.3%(68.2%, 90.9%)	93.9%(84.0%, 100.0%)
Sleep - Awake	50-60	0.973(0.939, 0.994)	0.991(0.978, 0.998)	65.9%(45.5%, 84.2%)	80.5%(70.3%, 90.9%)
Sleep - Awake	60-70	0.985(0.956, 1.000)	0.995(0.985, 1.000)	74.2%(54.5%, 88.7%)	88.7%(77.3%, 95.5%)
Sleep - Awake	>70	0.980(0.951, 1.000)	0.997(0.991, 1.000)	62.1%(40.9%, 77.3%)	91.3%(74.9%, 100.0%)
Sleep - N1	0-20	0.970 (0.911, 1.000)	0.986 (0.959, 1.000)	59.4% (36.4%, 84.2%)	76.0% (61.2%, 90.9%)
Sleep - N1	20-30	0.641 (0.473, 0.797)	0.178 (0.077, 0.306)	15.8% (9.1%, 41.1%)	15.2% (9.1%, 34.2%)
Sleep - N1	30-50	0.648(0.443, 0.790)	0.102(0.043, 0.181)	24.7%(22.7%, 31.8%)	24.5%(22.7%, 31.8%)
Sleep - N1	50-60	0.806(0.630, 0.906)	0.329(0.128, 0.592)	21.1%(18.2%, 36.4%)	20.5%(18.2%, 34.2%)
Sleep - N1	60-70	0.886(0.723, 0.998)	0.568(0.125, 0.985)	47.0%(22.7%, 81.8%)	46.5%(22.7%, 82.0%)
Sleep - N1	>70	0.610(0.441, 0.744)	0.077(0.029, 0.159)	40.6%(27.3%, 100.0%)	40.6%(27.3%, 100.0%)
Sleep - N2	0-20	0.597 (0.412, 0.761)	0.098 (0.030, 0.196)	33.0% (18.2%, 100.0%)	33.0% (18.2%, 100.0%)
Sleep - N2	20-30	0.948 (0.904, 0.986)	0.821 (0.678, 0.951)	20.8% (4.5%, 40.9%)	40.9% (4.5%, 93.3%)
Sleep - N2	30-50	0.993(0.981, 1.000)	0.963(0.889, 1.000)	32.5%(15.8%, 50.0%)	68.6%(31.8%, 100.0%)
Sleep - N2	50-60	0.991(0.973, 1.000)	0.956(0.883, 1.000)	56.1%(29.4%, 72.7%)	88.6%(74.9%, 95.5%)
Sleep - N2	60-70	0.989(0.971, 1.000)	0.936(0.822, 1.000)	39.0%(18.2%, 54.5%)	58.5%(36.4%, 95.5%)
Sleep - N2	>70	0.985(0.961, 1.000)	0.883(0.736, 1.000)	64.9%(36.1%, 86.4%)	77.9%(50.0%, 100.0%)
Spike	0-20	0.957 (0.915, 0.989)	0.764 (0.594, 0.904)	70.3% (43.1%, 90.9%)	75.4% (45.5%, 97.8%)
Spike	20-30	0.947 (0.897, 0.983)	0.808 (0.640, 0.940)	73.1% (36.4%, 95.5%)	76.5% (38.5%, 97.8%)
Spike	30-50	0.961(0.917, 0.990)	0.882(0.775, 0.948)	84.0%(61.2%, 97.8%)	86.0%(61.2%, 100.0%)
Spike	50-60	0.950(0.917, 0.973)	0.846(0.739, 0.918)	73.5%(52.2%, 90.9%)	76.1%(59.1%, 95.5%)
Spike	60-70	0.970(0.946, 0.993)	0.918(0.849, 0.972)	85.0%(72.7%, 95.5%)	94.3%(79.4%, 100.0%)
Spike	>70	0.966(0.949, 0.984)	0.904(0.860, 0.945)	75.5%(59.1%, 86.4%)	82.5%(68.2%, 93.3%)

1172

1173 (D) External SAI dataset

Task	Model	Age Group	PR AUC	PR EUC	ROC AUC	ROC EUC
Normal	Morgoth	0-10	1.000 (1.000, 1.000)	60.5% (36.4%, 72.7%)	1.000 (1.000, 1.000)	18.9% (0.0%, 27.3%)
Normal	Morgoth	10-30	0.998 (0.972, 1.000)	49.7% (17.7%, 71.4%)	0.998 (0.978, 1.000)	34.4% (7.1%, 50.0%)
Normal	Morgoth	30-50	1.000 (1.000, 1.000)	47.2% (22.5%, 63.6%)	1.000 (1.000, 1.000)	38.8% (4.3%, 54.5%)
Normal	Morgoth	50-70	0.928 (0.706, 1.000)	34.5% (0.0%, 90.9%)	0.959 (0.847, 1.000)	17.4% (9.1%, 27.3%)
Normal	Morgoth	>70	0.890 (0.533, 1.000)	15.9% (0.0%, 45.5%)	0.965 (0.874, 1.000)	25.6% (13.4%, 36.4%)
Normal	SCORE-AI	0-10	0.958 (0.750, 1.000)	44.8% (9.1%, 72.7%)	0.953 (0.800, 1.000)	13.1% (0.0%, 27.3%)
Normal	SCORE-AI	10-30	0.999 (0.991, 1.000)	58.1% (28.6%, 75.2%)	0.999 (0.989, 1.000)	39.4% (7.1%, 57.1%)
Normal	SCORE-AI	30-50	1.000 (1.000, 1.000)	47.2% (22.5%, 63.6%)	1.000 (1.000, 1.000)	38.8% (4.3%, 54.5%)
Normal	SCORE-AI	50-70	1.000 (1.000, 1.000)	67.4% (36.4%, 90.9%)	1.000 (1.000, 1.000)	23.4% (18.2%, 27.3%)
Normal	SCORE-AI	>70	1.000 (1.000, 1.000)	37.9% (18.2%, 45.5%)	1.000 (1.000, 1.000)	35.1% (22.5%, 45.5%)
Focal Slowing	Morgoth	0-10	0.983 (0.833, 1.000)	50.7% (27.3%, 72.7%)	0.989 (0.911, 1.000)	60.5% (36.4%, 90.9%)
Focal Slowing	Morgoth	10-30	0.948 (0.478, 1.000)	59.7% (32.0%, 85.7%)	0.975 (0.815, 1.000)	57.5% (35.7%, 71.4%)

Focal Slowing	Morgoth	30-50	0.989 (0.833, 1.000)	36.3% (9.1%, 78.4%)	0.998 (0.952, 1.000)	40.8% (18.2%, 87.0%)
Focal Slowing	Morgoth	50-70	1.000 (1.000, 1.000)	26.1% (9.1%, 36.4%)	1.000 (1.000, 1.000)	77.8% (45.5%, 90.9%)
Focal Slowing	Morgoth	>70	1.000 (1.000, 1.000)	34.5% (18.2%, 45.5%)	1.000 (1.000, 1.000)	38.5% (13.4%, 54.5%)
Focal Slowing	SCORE-AI	0-10	0.884 (0.361, 1.000)	45.2% (9.1%, 68.4%)	0.911 (0.615, 1.000)	42.9% (9.1%, 77.5%)
Focal Slowing	SCORE-AI	10-30	0.785 (0.417, 1.000)	42.9% (10.5%, 68.0%)	0.858 (0.578, 1.000)	43.5% (14.3%, 78.6%)
Focal Slowing	SCORE-AI	30-50	0.989 (0.833, 1.000)	36.3% (9.1%, 78.4%)	0.998 (0.952, 1.000)	40.8% (18.2%, 87.0%)
Focal Slowing	SCORE-AI	50-70	0.749 (0.434, 1.000)	11.9% (0.0%, 27.3%)	0.819 (0.519, 1.000)	30.1% (9.1%, 77.5%)
Focal Slowing	SCORE-AI	>70	0.934 (0.733, 1.000)	27.2% (9.1%, 45.5%)	0.946 (0.810, 1.000)	33.5% (9.1%, 54.5%)
Generalized Slowing	Morgoth	0-10	0.860 (0.333, 1.000)	74.8% (0.0%, 100.0%)	0.907 (0.538, 1.000)	70.5% (31.6%, 100.0%)
Generalized Slowing	Morgoth	10-30	0.966 (0.754, 1.000)	36.9% (21.4%, 50.0%)	0.993 (0.945, 1.000)	72.3% (50.0%, 92.9%)
Generalized Slowing	Morgoth	30-50	0.818 (0.333, 1.000)	28.3% (0.0%, 54.5%)	0.935 (0.800, 1.000)	36.5% (9.1%, 59.3%)
Generalized Slowing	Morgoth	50-70	0.943 (0.802, 1.000)	41.8% (18.2%, 54.5%)	0.934 (0.765, 1.000)	18.3% (0.0%, 54.5%)
Generalized Slowing	Morgoth	>70	1.000 (1.000, 1.000)	20.5% (0.0%, 27.3%)	1.000 (1.000, 1.000)	49.4% (36.4%, 59.3%)
Generalized Slowing	SCORE-AI	0-10	0.842 (0.273, 1.000)	72.0% (0.0%, 100.0%)	0.925 (0.667, 1.000)	76.9% (40.7%, 100.0%)
Generalized Slowing	SCORE-AI	10-30	0.912 (0.500, 1.000)	32.4% (14.3%, 50.0%)	0.982 (0.889, 1.000)	67.0% (42.9%, 85.7%)
Generalized Slowing	SCORE-AI	30-50	0.853 (0.500, 1.000)	29.4% (0.0%, 54.5%)	0.932 (0.743, 1.000)	31.3% (9.1%, 63.6%)
Generalized Slowing	SCORE-AI	50-70	0.837 (0.497, 1.000)	11.1% (0.0%, 54.5%)	0.910 (0.741, 1.000)	15.4% (0.0%, 36.4%)
Generalized Slowing	SCORE-AI	>70	0.885 (0.429, 1.000)	11.9% (0.0%, 18.2%)	0.863 (0.429, 1.000)	28.5% (0.0%, 54.5%)
Focal Spike	Morgoth	0-10	0.985 (0.750, 1.000)	42.9% (0.0%, 100.0%)	0.989 (0.848, 1.000)	59.0% (9.1%, 100.0%)
Focal Spike	Morgoth	10-30	0.938 (0.769, 1.000)	29.6% (7.1%, 57.1%)	0.961 (0.859, 1.000)	23.7% (7.1%, 50.0%)
Focal Spike	Morgoth	30-50	0.649 (0.250, 1.000)	70.9% (54.5%, 100.0%)	0.931 (0.812, 1.000)	85.7% (54.5%, 100.0%)
Focal Spike	Morgoth	50-70	0.804 (0.278, 1.000)	43.8% (18.2%, 77.5%)	0.873 (0.644, 1.000)	35.0% (9.1%, 72.7%)
Focal Spike	Morgoth	>70	1.000 (1.000, 1.000)	44.4% (0.0%, 100.0%)	1.000 (1.000, 1.000)	74.8% (36.4%, 100.0%)
Focal Spike	SCORE-AI	0-10	0.825 (0.429, 1.000)	30.9% (0.0%, 100.0%)	0.876 (0.556, 1.000)	57.2% (27.3%, 100.0%)
Focal Spike	SCORE-AI	10-30	0.969 (0.854, 1.000)	44.6% (10.5%, 78.6%)	0.978 (0.905, 1.000)	31.9% (7.1%, 57.1%)
Focal Spike	SCORE-AI	30-50	0.360 (0.125, 1.000)	70.9% (54.5%, 100.0%)	0.793 (0.562, 1.000)	77.7% (54.5%, 100.0%)
Focal Spike	SCORE-AI	50-70	0.830 (0.433, 1.000)	48.2% (0.0%, 90.9%)	0.896 (0.625, 1.000)	33.6% (0.0%, 68.4%)
Focal Spike	SCORE-AI	>70	1.000 (1.000, 1.000)	44.4% (0.0%, 100.0%)	1.000 (1.000, 1.000)	74.8% (36.4%, 100.0%)
Generalized Spike	Morgoth	0-10	1.000 (1.000, 1.000)	21.4% (0.0%, 100.0%)	1.000 (1.000, 1.000)	35.7% (0.0%, 100.0%)
Generalized Spike	Morgoth	10-30	0.871 (0.525, 1.000)	25.9% (0.0%, 50.0%)	0.971 (0.867, 1.000)	36.3% (0.0%, 78.6%)
Generalized Spike	Morgoth	30-50	1.000 (1.000, 1.000)	29.1% (9.1%, 100.0%)	1.000 (1.000, 1.000)	21.6% (9.1%, 100.0%)
Generalized Spike	Morgoth	50-70	1.000 (1.000, 1.000)	52.7% (27.3%, 100.0%)	1.000 (1.000, 1.000)	56.3% (27.3%, 100.0%)
Generalized Spike	Morgoth	>70	1.000 (1.000, 1.000)	39.5% (18.2%, 100.0%)	1.000 (1.000, 1.000)	39.5% (18.2%, 100.0%)
Generalized Spike	SCORE-AI	0-10	1.000 (1.000, 1.000)	21.4% (0.0%, 100.0%)	1.000 (1.000, 1.000)	35.7% (0.0%, 100.0%)
Generalized Spike	SCORE-AI	10-30	0.973 (0.808, 1.000)	33.7% (0.0%, 50.0%)	0.988 (0.906, 1.000)	34.8% (0.0%, 78.6%)
Generalized Spike	SCORE-AI	30-50	1.000 (1.000, 1.000)	29.1% (9.1%, 100.0%)	1.000 (1.000, 1.000)	21.6% (9.1%, 100.0%)
Generalized Spike	SCORE-AI	50-70	0.650 (0.250, 1.000)	52.7% (27.3%, 100.0%)	0.946 (0.833, 1.000)	54.6% (27.3%, 100.0%)
Generalized Spike	SCORE-AI	>70	1.000 (1.000, 1.000)	39.5% (18.2%, 100.0%)	1.000 (1.000, 1.000)	39.5% (18.2%, 100.0%)

1174

1175

1176 (E) External ON dataset

Task	Age Group	PR AUC	PR EUC	ROC AUC	ROC EUC
Focal Spike	0-10	0.952 (0.750, 1.000)	33.7% (7.1%, 50.0%)	0.987 (0.926, 1.000)	48.4% (14.3%, 71.4%)

Focal Spike	10-20	0.877 (0.452, 1.000)	22.0% (0.0%, 50.0%)	0.931 (0.740, 1.000)	22.9% (3.4%, 42.9%)
Focal Spike	20-40	0.879 (0.343, 1.000)	48.8% (0.0%, 100.0%)	0.965 (0.803, 1.000)	58.1% (0.0%, 100.0%)
Focal Spike	40-70	0.749 (0.440, 0.941)	9.9% (0.0%, 42.9%)	0.811 (0.553, 0.971)	3.3% (0.0%, 21.4%)
Focal Spike	>70	1.000 (1.000, 1.000)	51.1% (21.4%, 100.0%)	1.000 (1.000, 1.000)	61.7% (21.4%, 100.0%)
Generalized Spike	0-10	0.982 (0.920, 1.000)	25.7% (0.0%, 64.3%)	0.994 (0.967, 1.000)	34.4% (7.1%, 64.3%)
Generalized Spike	10-20	1.000 (1.000, 1.000)	25.4% (7.1%, 35.7%)	1.000 (1.000, 1.000)	9.0% (0.0%, 14.3%)
Generalized Spike	20-40	1.000 (1.000, 1.000)	46.6% (21.4%, 100.0%)	1.000 (1.000, 1.000)	53.1% (21.4%, 100.0%)
Generalized Spike	40-70	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)
Generalized Spike	>70	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)	1.000 (1.000, 1.000)	100.0% (100.0%, 100.0%)
Focal Slowing	0-10	0.882 (0.353, 1.000)	30.4% (7.1%, 100.0%)	0.983 (0.918, 1.000)	70.6% (35.7%, 100.0%)
Focal Slowing	10-20	0.717 (0.173, 1.000)	36.8% (14.3%, 100.0%)	0.896 (0.592, 1.000)	61.7% (28.6%, 100.0%)
Focal Slowing	20-40	1.000 (1.000, 1.000)	62.9% (42.9%, 100.0%)	1.000 (1.000, 1.000)	78.1% (42.9%, 100.0%)
Focal Slowing	40-70	0.799 (0.385, 1.000)	18.8% (0.0%, 35.7%)	0.891 (0.683, 1.000)	31.4% (7.1%, 78.6%)
Focal Slowing	>70	1.000 (1.000, 1.000)	55.4% (42.9%, 83.0%)	1.000 (1.000, 1.000)	73.4% (45.9%, 96.6%)
Generalized Slowing	0-10	0.606 (0.111, 1.000)	27.5% (3.4%, 100.0%)	0.847 (0.534, 1.000)	42.9% (7.1%, 100.0%)
Generalized Slowing	10-20	0.678 (0.298, 1.000)	49.7% (21.4%, 100.0%)	0.948 (0.833, 1.000)	54.9% (21.4%, 100.0%)
Generalized Slowing	20-40	0.707 (0.190, 1.000)	41.6% (10.5%, 100.0%)	0.767 (0.388, 1.000)	41.6% (7.1%, 100.0%)
Generalized Slowing	40-70	0.907 (0.721, 1.000)	45.6% (21.4%, 71.4%)	0.941 (0.812, 1.000)	43.9% (14.3%, 78.6%)
Generalized Slowing	>70	0.903 (0.732, 1.000)	53.2% (7.1%, 100.0%)	0.866 (0.666, 1.000)	24.4% (0.0%, 71.4%)

1177

1178 Table S8 Significance Test of Age Group Results
1179 (A) Internal IIIC-Test dataset

Model	Class	Metric	Average P-value	Proportion Significant
Kaggle Winner	GPD	AUC-ROC	0.43727418891386700	0
Kaggle Winner	GPD	AUC-PR	0.43727418891386700	0
Kaggle Winner	GPD	EUC-ROC	0.43727418891386700	0
Kaggle Winner	GPD	EUC-PR	0.43727418891386700	0
Kaggle Winner	GRDA	AUC-ROC	0.43727418891386700	0
Kaggle Winner	GRDA	AUC-PR	0.43727418891386700	0
Kaggle Winner	GRDA	EUC-ROC	0.43727418891386700	0
Kaggle Winner	GRDA	EUC-PR	0.43727418891386700	0
Kaggle Winner	LPD	AUC-ROC	0.44049328506521300	0
Kaggle Winner	LPD	AUC-PR	0.44049328506521300	0
Kaggle Winner	LPD	EUC-ROC	0.44049328506521300	0
Kaggle Winner	LPD	EUC-PR	0.44049328506521300	0
Kaggle Winner	LRDA	AUC-ROC	0.43727418891386700	0
Kaggle Winner	LRDA	AUC-PR	0.43727418891386700	0
Kaggle Winner	LRDA	EUC-ROC	0.43727418891386700	0
Kaggle Winner	LRDA	EUC-PR	0.43727418891386700	0
Kaggle Winner	Other	AUC-ROC	0.44049328506521300	0
Kaggle Winner	Other	AUC-PR	0.44049328506521300	0

Kaggle Winner	Other	EUC-ROC	0.44049328506521300	0
Kaggle Winner	Other	EUC-PR	0.44049328506521300	0
Kaggle Winner	Seizure	AUC-ROC	0.44049328506521300	0
Kaggle Winner	Seizure	AUC-PR	0.44049328506521300	0
Kaggle Winner	Seizure	EUC-ROC	0.44049328506521300	0
Kaggle Winner	Seizure	EUC-PR	0.44049328506521300	0
Morgoth	GPD	AUC-ROC	0.43727418891386700	0
Morgoth	GPD	AUC-PR	0.43727418891386700	0
Morgoth	GPD	EUC-ROC	0.43727418891386700	0
Morgoth	GPD	EUC-PR	0.43727418891386700	0
Morgoth	GRDA	AUC-ROC	0.43727418891386700	0
Morgoth	GRDA	AUC-PR	0.43727418891386700	0
Morgoth	GRDA	EUC-ROC	0.43727418891386700	0
Morgoth	GRDA	EUC-PR	0.43727418891386700	0
Morgoth	LPD	AUC-ROC	0.44049328506521300	0
Morgoth	LPD	AUC-PR	0.44049328506521300	0
Morgoth	LPD	EUC-ROC	0.44049328506521300	0
Morgoth	LPD	EUC-PR	0.44049328506521300	0
Morgoth	LRDA	AUC-ROC	0.43727418891386700	0
Morgoth	LRDA	AUC-PR	0.43727418891386700	0
Morgoth	LRDA	EUC-ROC	0.43727418891386700	0
Morgoth	LRDA	EUC-PR	0.43727418891386700	0
Morgoth	Other	AUC-ROC	0.44049328506521300	0
Morgoth	Other	AUC-PR	0.44049328506521300	0
Morgoth	Other	EUC-ROC	0.44049328506521300	0
Morgoth	Other	EUC-PR	0.44049328506521300	0
Morgoth	Seizure	AUC-ROC	0.44049328506521300	0
Morgoth	Seizure	AUC-PR	0.44049328506521300	0
Morgoth	Seizure	EUC-ROC	0.44049328506521300	0
Morgoth	Seizure	EUC-PR	0.44049328506521300	0
SPaRCNet	GPD	AUC-ROC	0.43727418891386700	0
SPaRCNet	GPD	AUC-PR	0.43727418891386700	0
SPaRCNet	GPD	EUC-ROC	0.43727418891386700	0
SPaRCNet	GPD	EUC-PR	0.43727418891386700	0
SPaRCNet	GRDA	AUC-ROC	0.43727418891386700	0
SPaRCNet	GRDA	AUC-PR	0.43727418891386700	0
SPaRCNet	GRDA	EUC-ROC	0.43727418891386700	0
SPaRCNet	GRDA	EUC-PR	0.43727418891386700	0
SPaRCNet	LPD	AUC-ROC	0.4404932850652120	0
SPaRCNet	LPD	AUC-PR	0.44049328506521300	0
SPaRCNet	LPD	EUC-ROC	0.44049328506521300	0
SPaRCNet	LPD	EUC-PR	0.44049328506521300	0
SPaRCNet	LRDA	AUC-ROC	0.43727418891386700	0

SPaRCNet	LRDA	AUC-PR	0.43727418891386700	0
SPaRCNet	LRDA	EUC-ROC	0.43727418891386700	0
SPaRCNet	LRDA	EUC-PR	0.43727418891386700	0
SPaRCNet	Other	AUC-ROC	0.44049328506521300	0
SPaRCNet	Other	AUC-PR	0.44049328506521300	0
SPaRCNet	Other	EUC-ROC	0.44049328506521300	0
SPaRCNet	Other	EUC-PR	0.44049328506521300	0
SPaRCNet	Seizure	AUC-ROC	0.44049328506521300	0
SPaRCNet	Seizure	AUC-PR	0.44049328506521300	0
SPaRCNet	Seizure	EUC-ROC	0.44049328506521300	0
SPaRCNet	Seizure	EUC-PR	0.44049328506521300	0

(B) Internal SN2-Test dataset

Model	Metric	Average P-value	Proportion Significant
Morgoth	PR AUC	1.0	0.0
Morgoth	PR EUC	0.5374741190774280	0.3055555555555560
Morgoth	ROC AUC	1.0	0.0
Morgoth	ROC EUC	1.0	0.0
SpikeNet2	PR AUC	1.0	0.0
SpikeNet2	PR EUC	0.7470533365933890	0.1388888888888890
SpikeNet2	ROC AUC	1.0	0.0
SpikeNet2	ROC EUC	0.7480075458324030	0.1388888888888890

(C) Internal MoE dataset

EEG Pattern	Metric	Significance Rate (%)	Mean p-value	Overall Avg Significant Rate (%)
Burst Suppression	ROC AUC	33.3	0.137	16.7
Burst Suppression	PR AUC	26.7	0.149	16.7
Burst Suppression	ROC EUC	6.7	0.181	16.7
Burst Suppression	PR EUC	0.0	0.2	16.7
Focal Slowing	ROC AUC	33.3	0.137	16.6
Focal Slowing	PR AUC	33.3	0.13	16.6
Focal Slowing	ROC EUC	0.0	0.187	16.6
Focal Slowing	PR EUC	0.0	0.18	16.6
Focal Spikes	ROC AUC	33.3	0.137	13.3
Focal Spikes	PR AUC	0.0	0.193	13.3
Focal Spikes	ROC EUC	13.3	0.18	13.3
Focal Spikes	PR EUC	6.7	0.183	13.3
Generalized Slowing	ROC AUC	33.3	0.117	51.6
Generalized Slowing	PR AUC	73.3	0.063	51.6
Generalized Slowing	ROC EUC	53.3	0.087	51.6
Generalized Slowing	PR EUC	46.7	0.09	51.6
Generalized Spikes	ROC AUC	20.0	0.154	10.0

Generalized Spikes	PR AUC	6.7	0.19	10.0
Generalized Spikes	ROC EUC	6.7	0.177	10.0
Generalized Spikes	PR EUC	6.7	0.183	10.0
IIIC - GPD	ROC AUC	0.0	0.2	1.7
IIIC - GPD	PR AUC	0.0	0.2	1.7
IIIC - GPD	ROC EUC	6.7	0.19	1.7
IIIC - GPD	PR EUC	0.0	0.2	1.7
IIIC - GRDA	ROC AUC	20.0	0.162	26.7
IIIC - GRDA	PR AUC	26.7	0.143	26.7
IIIC - GRDA	ROC EUC	26.7	0.135	26.7
IIIC - GRDA	PR EUC	33.3	0.147	26.7
IIIC - LPD	ROC AUC	6.7	0.19	5.0
IIIC - LPD	PR AUC	0.0	0.2	5.0
IIIC - LPD	ROC EUC	13.3	0.18	5.0
IIIC - LPD	PR EUC	0.0	0.2	5.0
IIIC - LRDA	ROC AUC	6.7	0.19	28.3
IIIC - LRDA	PR AUC	26.7	0.149	28.3
IIIC - LRDA	ROC EUC	46.7	0.106	28.3
IIIC - LRDA	PR EUC	33.3	0.145	28.3
IIIC - Seizure	ROC AUC	13.3	0.16	15.0
IIIC - Seizure	PR AUC	26.7	0.128	15.0
IIIC - Seizure	ROC EUC	0.0	0.2	15.0
IIIC - Seizure	PR EUC	20.0	0.147	15.0
Sleep - Awake	ROC AUC	0.0	0.193	8.3
Sleep - Awake	PR AUC	0.0	0.193	8.3
Sleep - Awake	ROC EUC	0.0	0.187	8.3
Sleep - Awake	PR EUC	33.3	0.141	8.3
Sleep - N1	ROC AUC	46.7	0.113	33.4
Sleep - N1	PR AUC	46.7	0.099	33.4
Sleep - N1	ROC EUC	20.0	0.155	33.4
Sleep - N1	PR EUC	20.0	0.135	33.4
Sleep - N2	ROC AUC	40.0	0.113	26.6
Sleep - N2	PR AUC	33.3	0.123	26.6
Sleep - N2	ROC EUC	20.0	0.161	26.6
Sleep - N2	PR EUC	13.3	0.164	26.6
Spike	ROC AUC	0.0	0.2	0.0
Spike	PR AUC	0.0	0.187	0.0
Spike	ROC EUC	0.0	0.2	0.0
Spike	PR EUC	0.0	0.193	0.0

1185
1186
1187 (D) External SAI dataset
1188

Model	Class	Metric	Average P-value	Proportion Significant
Morgoth	Normal	AUC-ROC	0.173	60.00%
Morgoth	Normal	AUC-PR	0.39	10.00%
Morgoth	Normal	EUC-ROC	0.237	40.00%
Morgoth	Normal	EUC-PR	0.445	0.00%
Morgoth	Focal Slowing	AUC-ROC	0.245	10.00%
Morgoth	Focal Slowing	AUC-PR	0.367	10.00%
Morgoth	Focal Slowing	EUC-ROC	0.189	20.00%
Morgoth	Focal Slowing	EUC-PR	0.223	30.00%
Morgoth	Generalized Slowing	AUC-ROC	0.234	30.00%
Morgoth	Generalized Slowing	AUC-PR	0.189	40.00%
Morgoth	Generalized Slowing	EUC-ROC	0.167	50.00%
Morgoth	Generalized Slowing	EUC-PR	0.145	60.00%
Morgoth	Focal Spike	AUC-ROC	0.089	60.00%
Morgoth	Focal Spike	AUC-PR	0.234	30.00%
Morgoth	Focal Spike	EUC-ROC	0.145	50.00%
Morgoth	Focal Spike	EUC-PR	0.067	70.00%
Morgoth	Generalized Spike	AUC-ROC	0.456	10.00%
Morgoth	Generalized Spike	AUC-PR	0.234	30.00%
Morgoth	Generalized Spike	EUC-ROC	0.567	0.00%
Morgoth	Generalized Spike	EUC-PR	0.345	20.00%
SCORE-AI	Normal	AUC-ROC	0.089	20.00%
SCORE-AI	Normal	AUC-PR	0.456	0.00%
SCORE-AI	Normal	EUC-ROC	0.124	10.00%
SCORE-AI	Normal	EUC-PR	0.278	0.00%
SCORE-AI	Focal Slowing	AUC-ROC	0.156	50.00%
SCORE-AI	Focal Slowing	AUC-PR	0.234	30.00%
SCORE-AI	Focal Slowing	EUC-ROC	0.178	40.00%
SCORE-AI	Focal Slowing	EUC-PR	0.567	0.00%
SCORE-AI	Generalized Slowing	AUC-ROC	0.567	0.00%
SCORE-AI	Generalized Slowing	AUC-PR	0.078	70.00%
SCORE-AI	Generalized Slowing	EUC-ROC	0.245	30.00%
SCORE-AI	Generalized Slowing	EUC-PR	0.089	60.00%
SCORE-AI	Focal Spike	AUC-ROC	0.034	80.00%
SCORE-AI	Focal Spike	AUC-PR	0.345	20.00%
SCORE-AI	Focal Spike	EUC-ROC	0.089	60.00%
SCORE-AI	Focal Spike	EUC-PR	0.123	50.00%
SCORE-AI	Generalized Spike	AUC-ROC	0.045	70.00%
SCORE-AI	Generalized Spike	AUC-PR	0.189	40.00%
SCORE-AI	Generalized Spike	EUC-ROC	0.234	30.00%
SCORE-AI	Generalized Spike	EUC-PR	0.156	50.00%

1189

1190 (E) External ON dataset

Class	Metric	Average P-value	Proportion Significant
Focal Slowing	AUC-ROC	0.067	60.00%
Focal Slowing	AUC-PR	0.089	60.00%
Focal Slowing	EUC-ROC	0.045	70.00%
Focal Slowing	EUC-PR	0.034	80.00%
Generalized Slowing	AUC-ROC	0.234	30.00%
Generalized Slowing	AUC-PR	0.067	70.00%
Generalized Slowing	EUC-ROC	0.345	20.00%
Generalized Slowing	EUC-PR	0.045	80.00%
Focal Spike	AUC-ROC	0.089	60.00%
Focal Spike	AUC-PR	0.123	50.00%
Focal Spike	EUC-ROC	0.167	40.00%
Focal Spike	EUC-PR	0.145	50.00%
Generalized Spike	AUC-ROC	0.078	60.00%
Generalized Spike	AUC-PR	0.234	30.00%
Generalized Spike	EUC-ROC	0.189	40.00%
Generalized Spike	EUC-PR	0.345	20.00%

12. Validation on Diverse International Cohorts

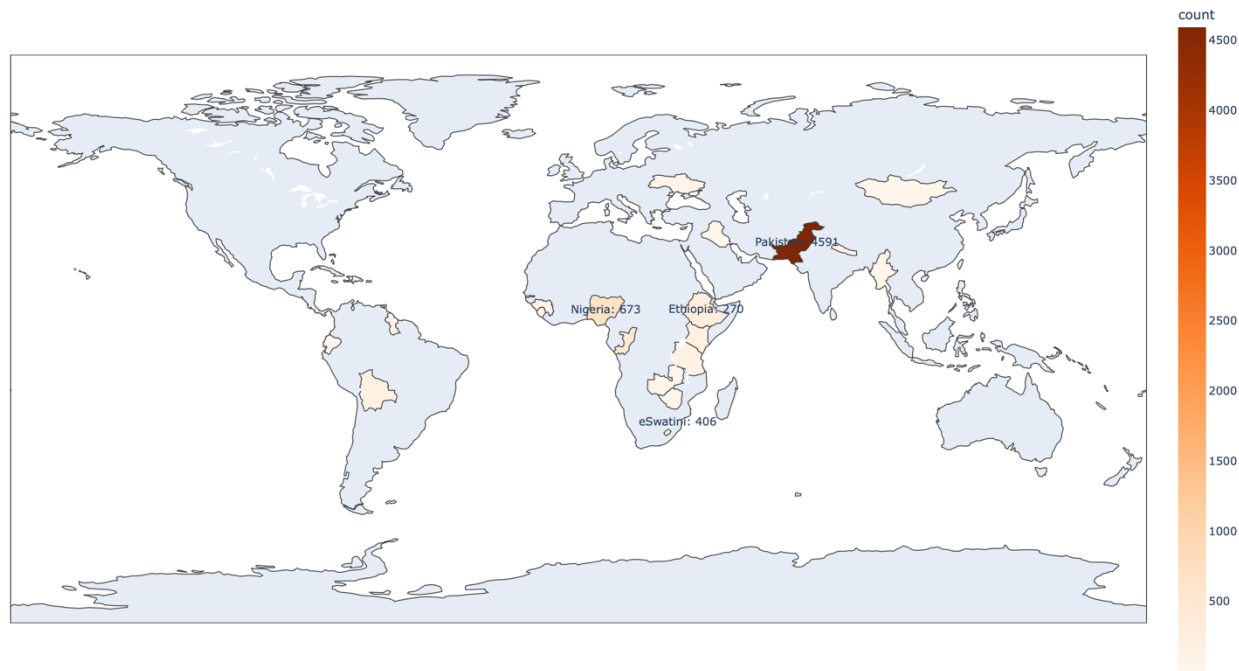
To evaluate the generalizability of our model, we conducted additional experiments on the TELEEEG dataset.

This dataset comprises 8,099 EEG recordings from individual participants collected across 23 countries and 42 clinical sites. Each entry includes demographic information such as sex and age at the time of visit (ranging from infancy to over 100 years old), as well as EEG-specific metadata including recording start time, duration, sampling rate, number of channels, and original file type. The cohort contains both pediatric and adult patients, with a roughly balanced distribution of male and female subjects. This diverse dataset supports downstream analyses of EEG patterns across a wide range of ages, geographies, and acquisition settings. Each EEG recording in the dataset is paired with a corresponding expert-written clinical report, typically authored by board-certified neurologists or epilepsy specialists. These reports contain free-text interpretations of the EEG findings.

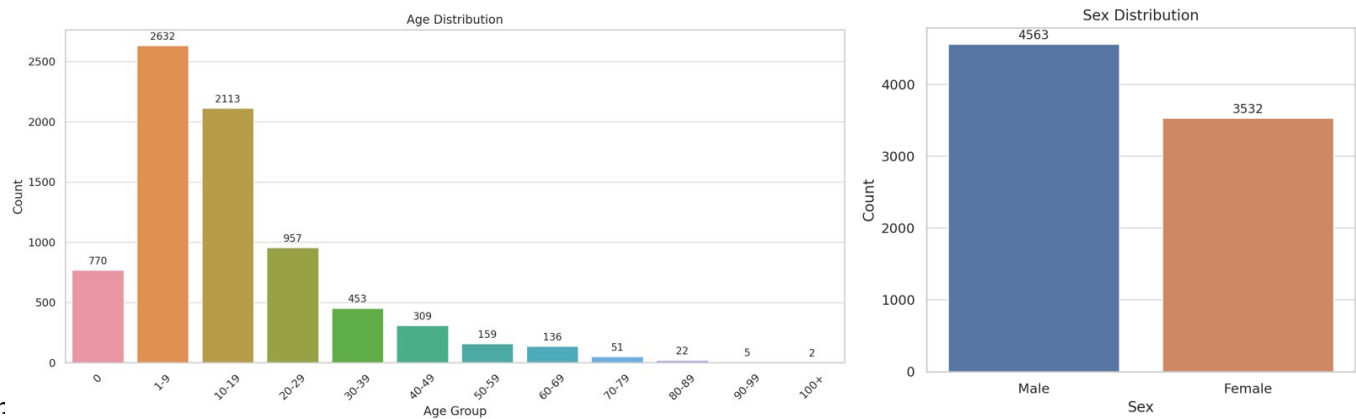
Unlike the primary dataset described in the main text, which consists predominantly of EEGs from Europe and North America, the majority of subjects in this dataset are from South America, Africa, the Middle East, and other under-resourced regions (Figure S24A).

This introduces several challenges for model validation and generalization. First, the demographic composition differs substantially, with greater ethnic and geographic diversity compared to the original training population. Second, there is a notably higher proportion of infants and young children (Figure S24B), whose EEG characteristics differ from those of older children and adults.

Third, many recordings exhibit incomplete electrode coverage, often lacking the full set of 19 channels used in Morgoth’s training. Additionally, sampling rates vary widely across recordings, further complicating data harmonization. While expert-written reports are available, they are not always comprehensive, and heterogeneous reporting styles make automated label extraction difficult and potentially unreliable. These limitations highlight the importance of caution when using this dataset for validation purposes.



(A) Distribution of Subjects by Country



(B) Age and Sex Distribution

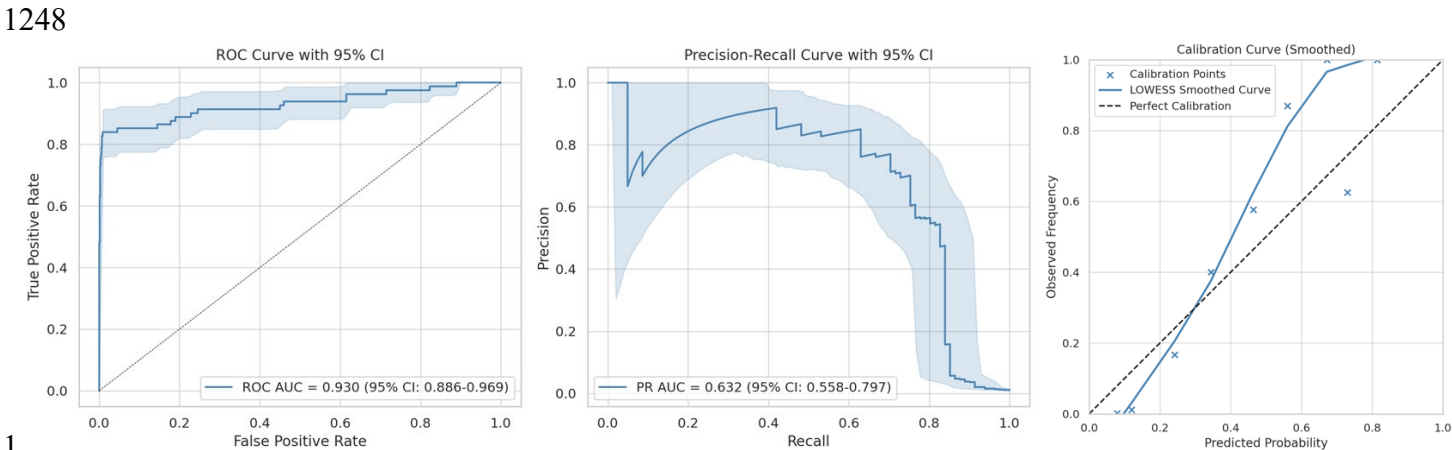
Figure S24. Distribution

We used both regular expressions and Medical LLaMA3 to extract EEG-level labels from the free-text reports. Only labels for which both methods produced consistent results were retained as valid annotations. These labels reflect overall findings at the EEG level (e.g., presence of spikes, slowing), and thus can only be used to evaluate EEG-level classification tasks; event-level assessments are not possible.

Here we evaluate EEG-level detection of normal activity, burst suppression, spikes, seizures, focal slowing, generalized slowing, and sleep stage N1, N2. Other features, such as LPD, GPD, LRDA, GRDA, sleep stage N3, and REM, were not mentioned in the reports.

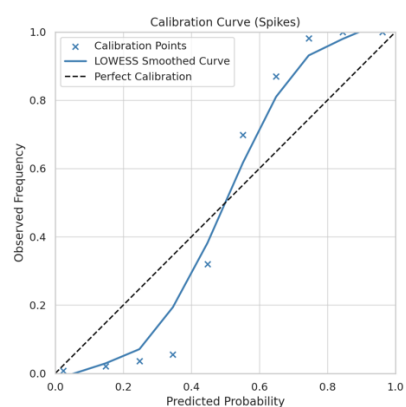
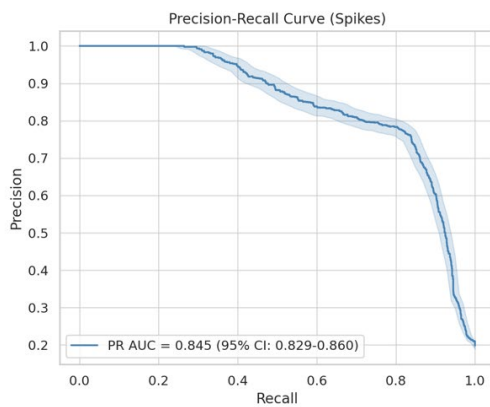
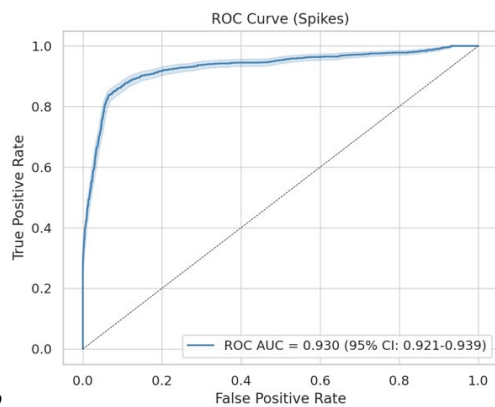
As shown in Figure S25, overall performance was reasonable, with ROC AUC values exceeding 0.9 for all tasks except sleep. However, compared to other test sets, the model showed a significant drop in performance for spike detection, seizure detection, and sleep stage N2, as evidenced by non-overlapping confidence intervals.

The performance decline can be attributed to several factors. From a model perspective, one key limitation is the absence of certain EEG channels in the external dataset, which may hinder the model's ability to fully interpret patterns. On the training side, the dataset may lack demographic diversity—particularly underrepresented are infants and populations from low-resource regions—potentially affecting generalization. Moreover, the reliability of the external labels is inherently limited: some clinically relevant findings may not be explicitly documented in the EEG reports, and many labels lack validation by multiple experts. These issues collectively reduce both the fidelity of the ground truth and the alignment between training and evaluation conditions.



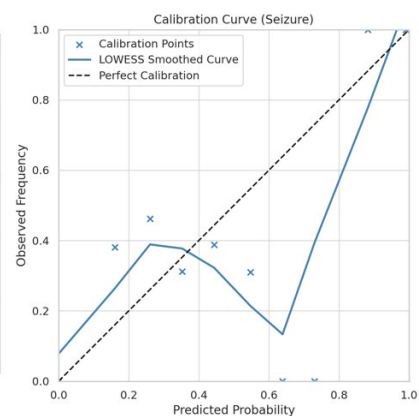
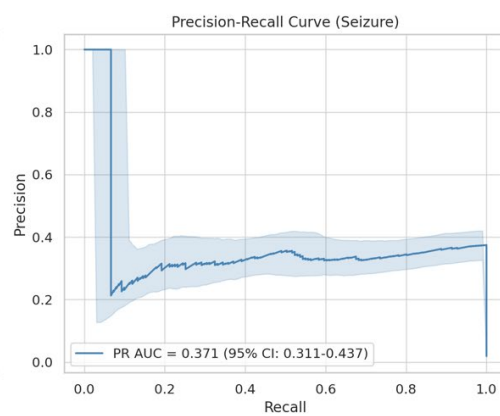
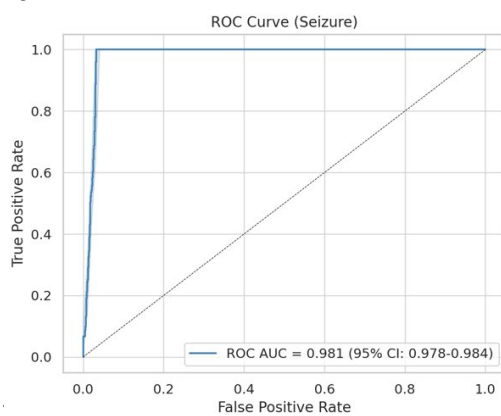
Burst suppression

12
1253
1254



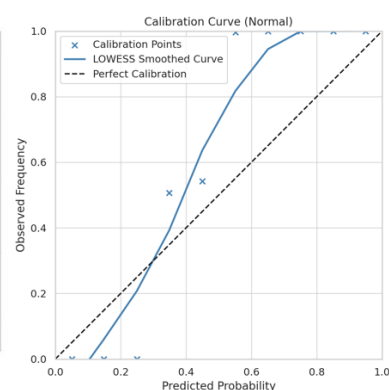
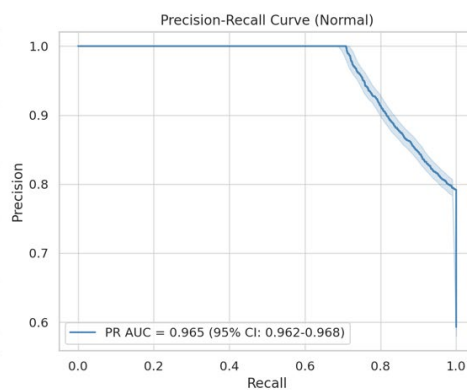
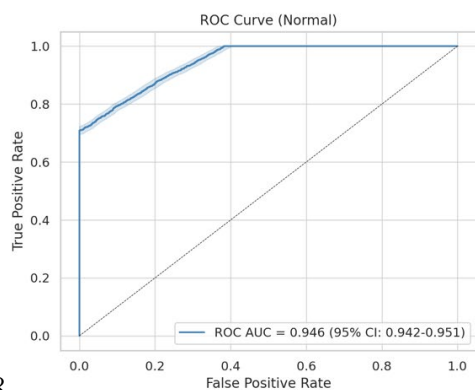
Spike

1
1256
1257

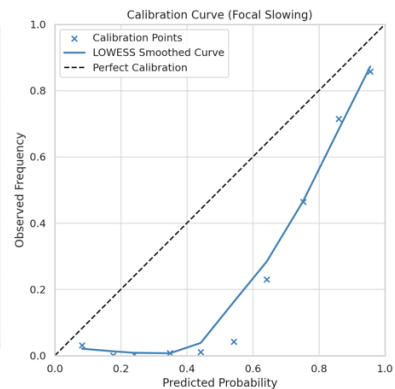
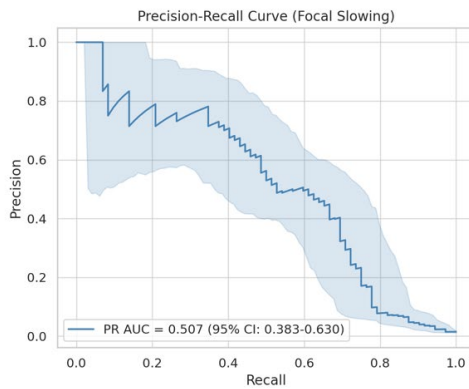
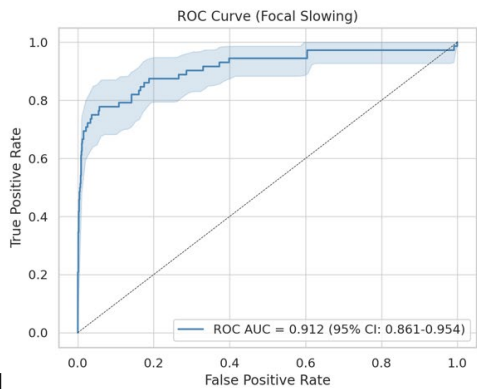


Seizure

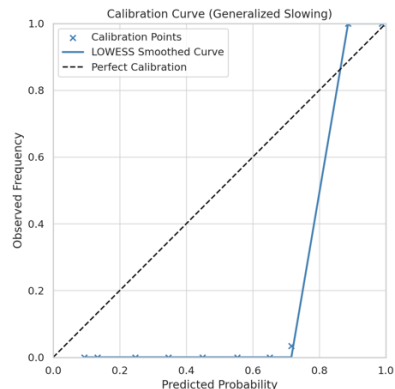
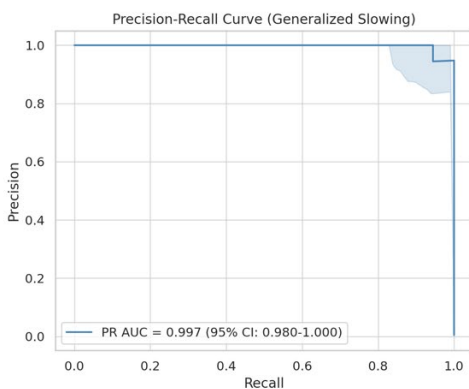
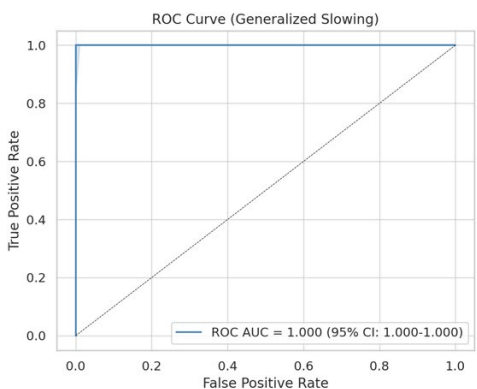
1258
1259
1260



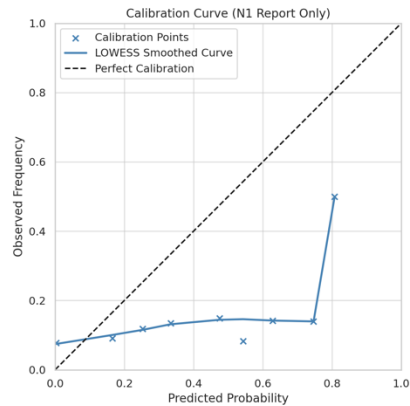
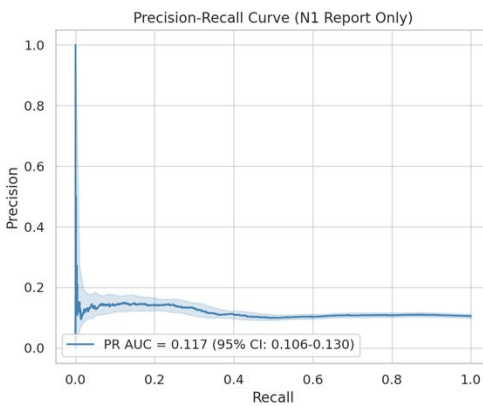
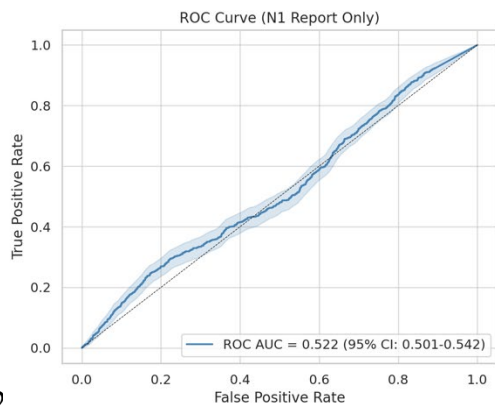
Normal vs Abnormal



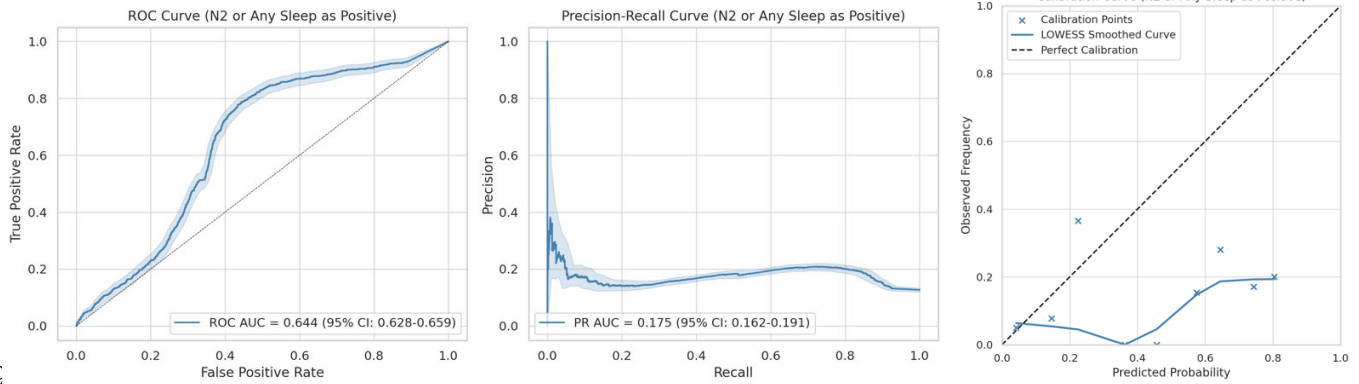
Focal Slowing



Generalized Slowing



Sleep N1



Sleep N2

Figure S25. Performance Results on External TELEEEG Datasets

13. Test on Missing Channels Datasets

To evaluate the model's robustness under missing-channel scenarios, we categorized the 19 standard 10–20 EEG input channels into seven anatomical groups and simulated corresponding channel dropout conditions. These seven groups include:

- Fp-missing: loss of frontopolar channels (Fp1, Fp2),
- F-missing: loss of frontal channels (F3, F4, F7, F8),
- C-missing: loss of central channels (C3, C4, Cz),
- P-missing: loss of parietal channels (P3, P4, Pz),
- O-missing: loss of occipital channels (O1, O2),
- T-missing: loss of temporal channels (T3/T7, T4/T8, T5/P7, T6/P8),
- Z-missing: loss of midline electrodes (Fz, Cz, Pz).

MORGOTH can handle EEG recordings with missing channels by automatically imputing them using the mean of the available channels before feeding the data into the model. This evaluation allows us to assess the generalizability and resilience of the model across diverse missing-channel conditions.

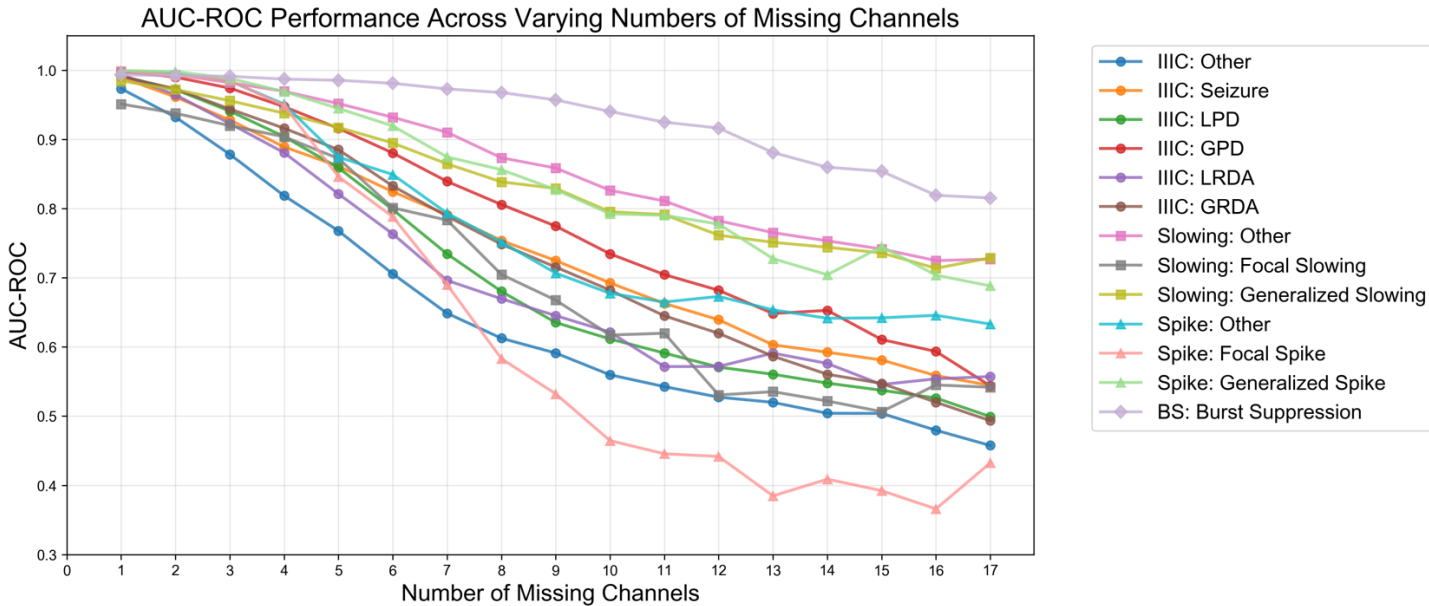
We evaluated model performance on the IIIC-Test and MoE datasets for the following findings: seizure, LPD, GPD, LRDA, GRDA, burst suppression, focal slowing, generalized slowing, focal spikes, and generalized spikes. As shown in Figure S26, the experiments are based on both the removal of varying numbers of channels and the removal of specific types of channels.

Under the removal of varying numbers of channels conditions, we removed channels randomly. As shown in the Figure S26A, increasing the number of missing EEG channels leads to a gradual decline in AUC-ROC across all classification tasks. Detection of focal spikes is most affected,

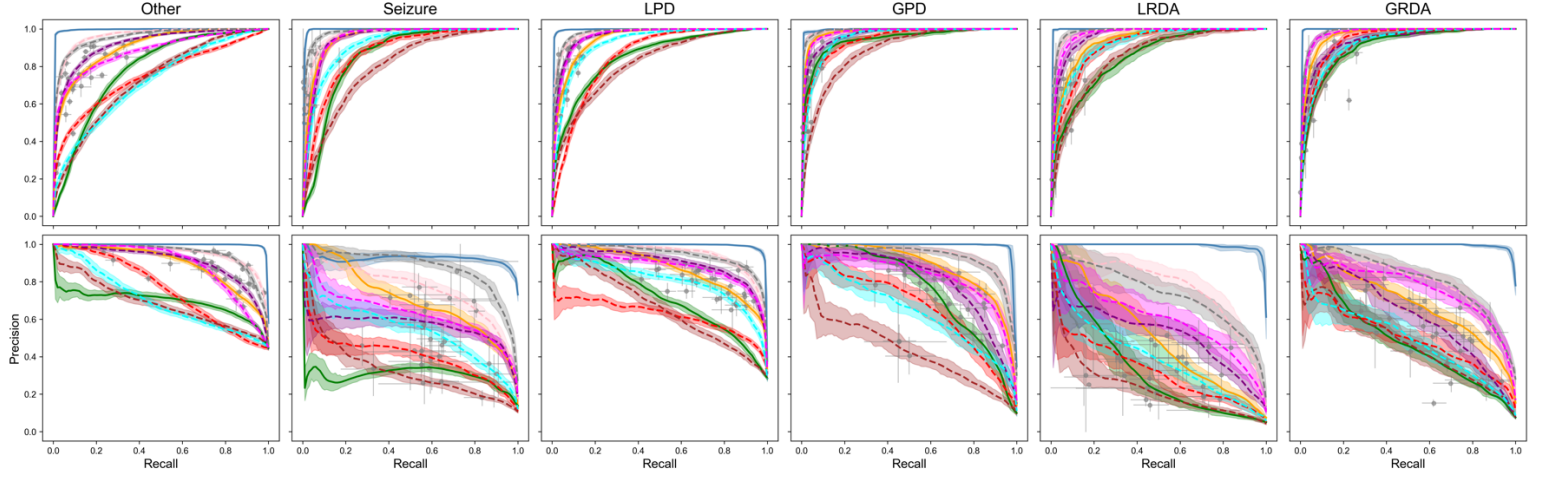
with performance dropping sharply after approximately 6–8 missing channels—indicating a clear inflection point beyond which detection becomes unreliable. Similarly, focal and generalized slowing exhibit steep declines beginning around 7–9 missing channels. In contrast, burst suppression remains highly robust, with minimal degradation even as up to 17 channels are removed, and no sharp inflection point is observed. Other features such as GRDA and LRDA show moderate declines, with more gradual inflection points around 10 missing channels. These results suggest that spatially localized features like spikes and slowing are highly sensitive to partial channel loss, especially beyond specific thresholds, while globally distributed patterns such as burst suppression are more resilient and less dependent on full spatial coverage.

Under the removal of specific types of channels conditions, MORGOTH consistently outperformed SPaRCNet, except in the cases where F- or Z-channels were missing. When P- or C-channels were missing, MORGOTH still outperformed the Kaggle Winner model. In all missing-channel scenarios, MORGOTH consistently outperformed the burst-suppression baseline.

For specific feature detection tasks, missing P channels had the least impact across all tasks. Missing F channels had the greatest impact on performance. Channel missing had the least impact on burst suppression detection. However, for features with strong spatial localization—such as LPD, GPD, focal slowing, and generalized slowing—the performance was more affected. For spatially localized features, replacing missing channels with the mean of other channels tends to “wash out” the distinctive abnormal patterns. These features rely heavily on detecting localized deviations from baseline activity. Once replaced with averaged values—often close to zero or normal background—the distinguishing signal may disappear entirely, making detection much harder for the model.



(A) AUC-ROC per Class Across Varying Numbers of Missing Channels



Morgoth:
 ROC AUC: Other=1.00 Seizure=0.99 LPD=1.00 GPD=1.00 LRDA=1.00 GRDA=1.00
 ROC EUC: Other=100% Seizure=92% LPD=100% GPD=100% LRDA=100% GRDA=100%
 PR AUC: Other=1.00 Seizure=0.92 LPD=0.99 GPD=0.99 LRDA=0.99 GRDA=0.99
 PR EUC: Other=100% Seizure=96% LPD=100% GPD=100% LRDA=100% GRDA=100%

Kaggle Winner:
 ROC AUC: Other=0.91 Seizure=0.95 LPD=0.95 GPD=0.98 LRDA=0.94 GRDA=0.97
 ROC EUC: Other=50% Seizure=32% LPD=36% GPD=98% LRDA=68% GRDA=70%
 PR AUC: Other=0.89 Seizure=0.70 LPD=0.87 GPD=0.88 LRDA=0.50 GRDA=0.69
 PR EUC: Other=26% Seizure=74% LPD=53% GPD=92% LRDA=73% GRDA=73%

SPARCNet:
 ROC AUC: Other=0.78 Seizure=0.86 LPD=0.85 GPD=0.94 LRDA=0.85 GRDA=0.91
 ROC EUC: Other=1% Seizure=0% LPD=1% GPD=52% LRDA=25% GRDA=19%
 PR AUC: Other=0.69 Seizure=0.30 LPD=0.71 GPD=0.76 LRDA=0.33 GRDA=0.48
 PR EUC: Other=0% Seizure=18% LPD=1% GPD=43% LRDA=31% GRDA=16%

missing xZ:
 ROC AUC: Other=0.75 Seizure=0.87 LPD=0.85 GPD=0.94 LRDA=0.91 GRDA=0.94
 ROC EUC: Other=9% Seizure=1% LPD=0% GPD=26% LRDA=32% GRDA=28%
 PR AUC: Other=0.76 Seizure=0.40 LPD=0.62 GPD=0.66 LRDA=0.33 GRDA=0.50
 PR EUC: Other=6% Seizure=30% LPD=0% GPD=16% LRDA=43% GRDA=27%

missing Fpx:
 ROC AUC: Other=0.92 Seizure=0.95 LPD=0.96 GPD=0.97 LRDA=0.96 GRDA=0.94
 ROC EUC: Other=64% Seizure=12% LPD=63% GPD=78% LRDA=85% GRDA=54%
 PR AUC: Other=0.91 Seizure=0.56 LPD=0.90 GPD=0.81 LRDA=0.51 GRDA=0.62
 PR EUC: Other=47% Seizure=66% LPD=87% GPD=69% LRDA=88% GRDA=53%

missing Fx:
 ROC AUC: Other=0.73 Seizure=0.80 LPD=0.82 GPD=0.88 LRDA=0.85 GRDA=0.91
 ROC EUC: Other=0% Seizure=0% LPD=0% GPD=1% LRDA=13% GRDA=20%
 PR AUC: Other=0.68 Seizure=0.33 LPD=0.66 GPD=0.43 LRDA=0.24 GRDA=0.47
 PR EUC: Other=0% Seizure=9% LPD=0% GPD=5% LRDA=18% GRDA=18%

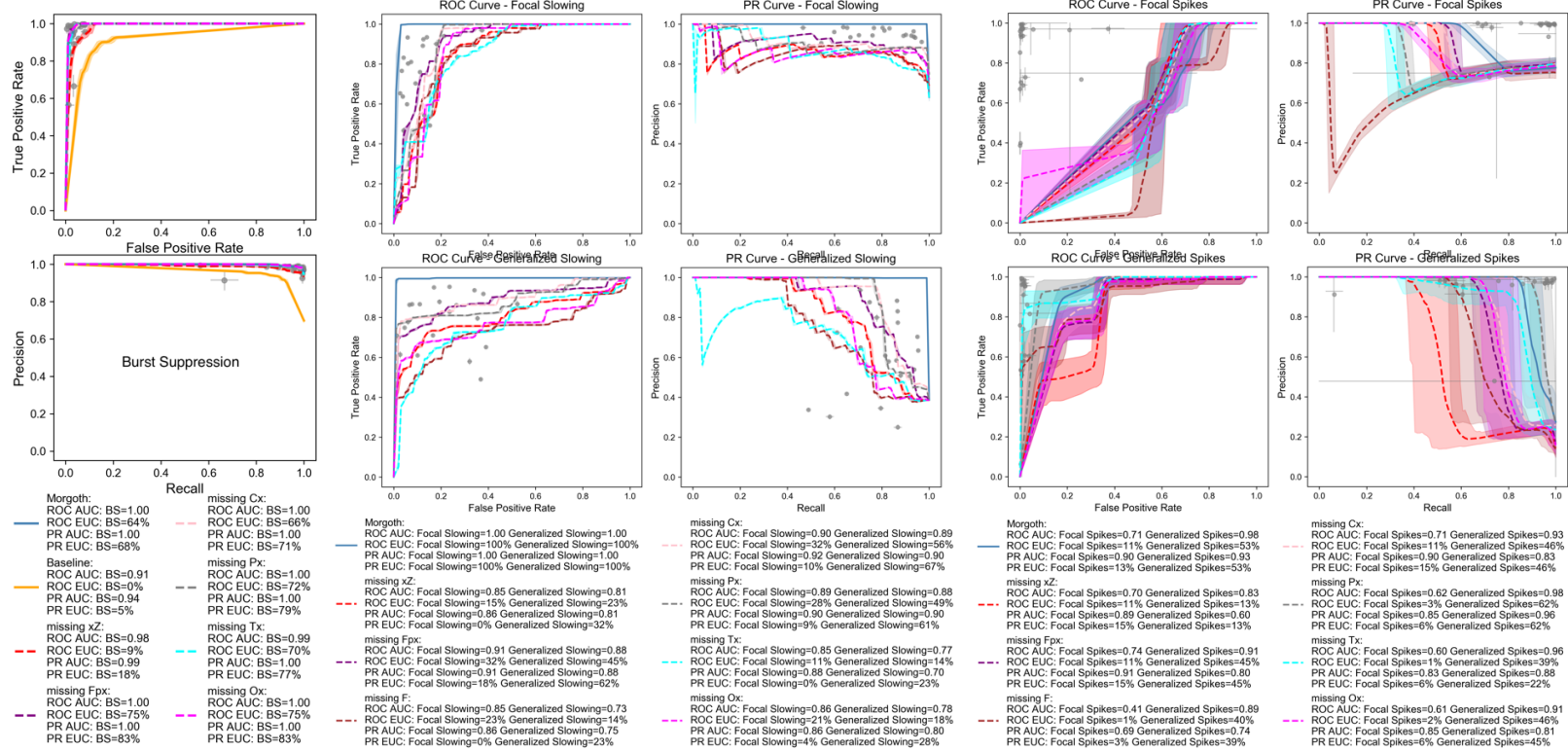
missing Cx:
 ROC AUC: Other=0.96 Seizure=0.97 LPD=0.98 GPD=0.99 LRDA=0.98 GRDA=0.98
 ROC EUC: Other=98% Seizure=54% LPD=94% GPD=99% LRDA=98% GRDA=90%
 PR AUC: Other=0.96 Seizure=0.76 LPD=0.95 GPD=0.94 LRDA=0.77 GRDA=0.79
 PR EUC: Other=63% Seizure=88% LPD=96% GPD=99% LRDA=99% GRDA=94%

missing Px:
 ROC AUC: Other=0.96 Seizure=0.98 LPD=0.98 GPD=0.99 LRDA=0.99 GRDA=0.98
 ROC EUC: Other=98% Seizure=85% LPD=94% GPD=99% LRDA=98% GRDA=93%
 PR AUC: Other=0.95 Seizure=0.88 LPD=0.95 GPD=0.94 LRDA=0.73 GRDA=0.82
 PR EUC: Other=81% Seizure=95% LPD=97% GPD=99% LRDA=99% GRDA=95%

missing Tx:
 ROC AUC: Other=0.72 Seizure=0.92 LPD=0.92 GPD=0.95 LRDA=0.92 GRDA=0.92
 ROC EUC: Other=0% Seizure=8% LPD=16% GPD=19% LRDA=49% GRDA=24%
 PR AUC: Other=0.70 Seizure=0.54 LPD=0.81 GPD=0.64 LRDA=0.37 GRDA=0.49
 PR EUC: Other=0% Seizure=54% LPD=28% GPD=13% LRDA=59% GRDA=23%

missing Ox:
 ROC AUC: Other=0.88 Seizure=0.95 LPD=0.96 GPD=0.98 LRDA=0.97 GRDA=0.98
 ROC EUC: Other=41% Seizure=21% LPD=50% GPD=93% LRDA=91% GRDA=86%
 PR AUC: Other=0.89 Seizure=0.63 LPD=0.88 GPD=0.83 LRDA=0.59 GRDA=0.76
 PR EUC: Other=27% Seizure=69% LPD=81% GPD=80% LRDA=94% GRDA=91%

1324 (B) The performance of the model and baselines on the IIIC-Test dataset with missing channels.
 1325



Morgoth:
 ROC AUC: BS=1.00
 ROC EUC: BS=64%
 PR AUC: BS=1.00
 PR EUC: BS=68%

Baseline:
 ROC AUC: BS=0.91
 ROC EUC: BS=0%
 PR AUC: BS=0.94
 PR EUC: BS=5%

missing xZ:
 ROC AUC: BS=0.98
 ROC EUC: BS=9%
 PR AUC: BS=0.99
 PR EUC: BS=16%

missing Fpx:
 ROC AUC: BS=1.00
 ROC EUC: BS=75%
 PR AUC: BS=1.00
 PR EUC: BS=83%

missing Cx:
 ROC AUC: BS=1.00
 ROC EUC: BS=86%
 PR AUC: BS=1.00
 PR EUC: BS=71%

missing Px:
 ROC AUC: BS=1.00
 ROC EUC: BS=72%
 PR AUC: BS=1.00
 PR EUC: BS=79%

missing Tx:
 ROC AUC: BS=0.99
 ROC EUC: BS=70%
 PR AUC: BS=1.00
 PR EUC: BS=77%

missing Ox:
 ROC AUC: BS=1.00
 ROC EUC: BS=75%
 PR AUC: BS=1.00
 PR EUC: BS=83%

Morgoth:
 ROC AUC: Focal Slowing=1.00 Generalized Slowing=1.00
 ROC EUC: Focal Slowing=100% Generalized Slowing=100%
 PR AUC: Focal Slowing=1.00 Generalized Slowing=1.00
 PR EUC: Focal Slowing=100% Generalized Slowing=100%

missing xZ:
 ROC AUC: Focal Slowing=0.85 Generalized Slowing=0.81
 ROC EUC: Focal Slowing=15% Generalized Slowing=23%
 PR AUC: Focal Slowing=0.86 Generalized Slowing=0.81
 PR EUC: Focal Slowing=0% Generalized Slowing=32%

missing Fpx:
 ROC AUC: Focal Slowing=0.91 Generalized Slowing=0.88
 ROC EUC: Focal Slowing=32% Generalized Slowing=46%
 PR AUC: Focal Slowing=0.91 Generalized Slowing=0.88
 PR EUC: Focal Slowing=18% Generalized Slowing=62%

missing Fx:
 ROC AUC: Focal Slowing=0.85 Generalized Slowing=0.73
 ROC EUC: Focal Slowing=23% Generalized Slowing=14%
 PR AUC: Focal Slowing=0.86 Generalized Slowing=0.75
 PR EUC: Focal Slowing=0% Generalized Slowing=23%

missing Cx:
 ROC AUC: Focal Slowing=0.90 Generalized Slowing=0.89
 ROC EUC: Focal Slowing=32% Generalized Slowing=56%
 PR AUC: Focal Slowing=0.92 Generalized Slowing=0.90
 PR EUC: Focal Slowing=10% Generalized Slowing=67%

missing Px:
 ROC AUC: Focal Slowing=0.89 Generalized Slowing=0.88
 ROC EUC: Focal Slowing=28% Generalized Slowing=49%
 PR AUC: Focal Slowing=0.90 Generalized Slowing=0.90
 PR EUC: Focal Slowing=0% Generalized Slowing=61%

missing Tx:
 ROC AUC: Focal Slowing=0.85 Generalized Slowing=0.77
 ROC EUC: Focal Slowing=11% Generalized Slowing=14%
 PR AUC: Focal Slowing=0.88 Generalized Slowing=0.70
 PR EUC: Focal Slowing=0% Generalized Slowing=23%

missing Ox:
 ROC AUC: Focal Slowing=0.86 Generalized Slowing=0.78
 ROC EUC: Focal Slowing=21% Generalized Slowing=18%
 PR AUC: Focal Slowing=0.86 Generalized Slowing=0.80
 PR EUC: Focal Slowing=4% Generalized Slowing=28%

Morgoth:
 ROC AUC: Focal Spikes=0.98 Generalized Spikes=0.98
 ROC EUC: Focal Spikes=11% Generalized Spikes=53%
 PR AUC: Focal Spikes=0.90 Generalized Spikes=0.93
 PR EUC: Focal Spikes=13% Generalized Spikes=53%

missing xZ:
 ROC AUC: Focal Spikes=0.70 Generalized Spikes=0.83
 ROC EUC: Focal Spikes=11% Generalized Spikes=13%
 PR AUC: Focal Spikes=0.89 Generalized Spikes=0.80
 PR EUC: Focal Spikes=15% Generalized Spikes=13%

missing Fpx:
 ROC AUC: Focal Spikes=0.74 Generalized Spikes=0.91
 ROC EUC: Focal Spikes=11% Generalized Spikes=45%
 PR AUC: Focal Spikes=0.89 Generalized Spikes=0.74
 PR EUC: Focal Spikes=15% Generalized Spikes=39%

missing Fx:
 ROC AUC: Focal Spikes=0.69 Generalized Spikes=0.74
 ROC EUC: Focal Spikes=11% Generalized Spikes=40%
 PR AUC: Focal Spikes=0.69 Generalized Spikes=0.74
 PR EUC: Focal Spikes=3% Generalized Spikes=39%

missing Cx:
 ROC AUC: Focal Spikes=0.71 Generalized Spikes=0.93
 ROC EUC: Focal Spikes=11% Generalized Spikes=46%
 PR AUC: Focal Spikes=0.90 Generalized Spikes=0.83
 PR EUC: Focal Spikes=15% Generalized Spikes=46%

missing Px:
 ROC AUC: Focal Spikes=0.62 Generalized Spikes=0.98
 ROC EUC: Focal Spikes=3% Generalized Spikes=62%
 PR AUC: Focal Spikes=0.90 Generalized Spikes=0.96
 PR EUC: Focal Spikes=6% Generalized Spikes=62%

missing Tx:
 ROC AUC: Focal Spikes=0.60 Generalized Spikes=0.96
 ROC EUC: Focal Spikes=2% Generalized Spikes=39%
 PR AUC: Focal Spikes=0.83 Generalized Spikes=0.88
 PR EUC: Focal Spikes=6% Generalized Spikes=22%

missing Ox:
 ROC AUC: Focal Spikes=0.61 Generalized Spikes=0.91
 ROC EUC: Focal Spikes=2% Generalized Spikes=46%
 PR AUC: Focal Spikes=0.85 Generalized Spikes=0.81
 PR EUC: Focal Spikes=6% Generalized Spikes=45%

1327 (C) The performance of the model and baselines on the MoE dataset with missing channels.
 1328

References

1. Zafar, S., Loddenkemper, T., Lee, J. W., et al. (2024). Harvard Electroencephalography Database (version 3.0). Brain Data Science Platform. <https://doi.org/10.60508/xbsx-vr76>
2. Westover, M. B., Moura Junior, V., Thomas, R., et al. (2023). The Human Sleep Project (version 2.0). Brain Data Science Platform. <https://doi.org/10.60508/qjbv-hg78>
3. Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., Daneshjou, R., Frankle, J., Liang, P., Carbin, M., Manning, C. D. (2024). BioMedLM: A 2.7B Parameter Language Model Trained on Biomedical Text. CoRR, abs/2403.18421.
4. Westover, M. B., Shafi, M. M., Ching, S., Chemali, J. J., Purdon, P. L., Cash, S. S., Brown, E. N. (2013). Real-time segmentation of burst suppression patterns in critical care EEG monitoring. *Journal of Neuroscience Methods*, 219(1), 131–141. <https://doi.org/10.1016/j.jneumeth.2013.07.003>
5. Jing, J., Sun, H., Kim, J. A., et al. (2019). Development of expert-level automated detection of epileptiform discharges during electroencephalogram interpretation. *JAMA Neurology*. <https://doi.org/10.1001/jamaneurol.2019.3485>
6. Jing, J., Ge, W., et al. (2023). Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17), e1750–e1762. <https://doi.org/10.1212/WNL.0000000000207127>
7. Jing, J., Ge, W., Struck, A. F., Fernandes, M. B., et al. (2023). Interrater reliability of expert electroencephalographers identifying seizures and rhythmic and periodic patterns in EEGs. *Neurology*, 100(17), e1737–e1749.
8. Jing, J., Herlopian, A., Karakis, I., et al. (2020). Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. *JAMA Neurology*, 77(1), 49–57. <https://doi.org/10.1001/jamaneurol.2019.3531>
9. Amorim, E., Zheng, W., Ghassemi, M., et al. (2023). I-CARE: International Cardiac Arrest REsearch consortium Electroencephalography Database (version 1.0). Brain Data Science Platform. <https://doi.org/10.60508/xwye-v214>
10. Amorim, E., Zheng, W., Ghassemi, M., et al. (2023). The International Cardiac Arrest Research (I-CARE) Consortium Electroencephalography Database. *Critical Care Medicine*.
11. Tveit, J., Aurlen, H., Plis, S., et al. (2023). Automated interpretation of clinical electroencephalograms using artificial intelligence. *JAMA Neurology*, 80(8), 805–812. <https://doi.org/10.1001/jamaneurol.2023.1645>
12. Buckwalter, G., Chhin, S., Rahman, S., Obeid, I., Picone, J. (2021). Recent advances in the TUH EEG Corpus: Improving the interrater agreement for artifacts and epileptiform events. In 2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB) (pp. 1–3). <https://doi.org/10.1109/SPMB52430.2021.9672301>

13. West, L. C., Summers, M., Tang, S., Hirt, L., Halpern, C. H., Maroni, D., Das, R., Gliske, S. V., Abosch, A., Kushida, C. A., Thompson, J. A. (2023). Evaluation of consensus sleep stage scoring of dysregulated sleep in Parkinson's disease. *Sleep Medicine*, 107, 236–242. <https://doi.org/10.1016/j.sleep.2023.04.031>
14. O'Reilly, C., Gosselin, N., Carrier, J., Nielsen, T. (2015). Montreal archive of sleep studies: An open-access resource for instrument benchmarking and exploratory research. *Journal of Sleep Research*, 23(6), 628–635.
15. Peh, W., Yao, Y., Dauwels, J. (2022). Transformer convolutional neural networks for automated artifact detection in scalp EEG. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 3599–3602).
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
17. Ali, M., Fromm, M., Thellmann, K., et al. (2024). Tokenizer choice for LLM training: Negligible or crucial? *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (Findings)*, 3907–3924.
18. Jiang W., Zhao L., Lu B. (2024). Large brain model for learning generic representations with tremendous EEG data in BCI. *International Conference on Learning Representations (ICLR)*.
19. Peng, Z., Dong, L., Bao, H., Ye, Q., Wei, F. (2022). BEiT v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. *CoRR*, abs/2208.06366.
20. Van Den Oord, A., Vinyals, O., et al. (2017). Neural Discrete Representation Learning. *Advances in Neural Information Processing Systems*, 30, 6306–6315.
21. Wu, D., Li, S., Yang, J., Sawan, M. (2022). Neuro2vec: Masked Fourier Spectrum Prediction for Neurophysiological Representation Learning. *arXiv preprint, arXiv:2204.12440*.
22. Sun, C., Li, H., Li, Y., Hong, S. (2024). TEST: Text Prototype Aligned Embedding to Activate LLM's Ability for Time Series. *International Conference on Learning Representations (ICLR)*.
23. Meng, Q., Qian, H., Liu, Y., Cui, L., Xu, Y., Shen, Z. (2023). MHCCL: Masked Hierarchical Cluster-Wise Contrastive Learning for Multivariate Time Series. *AAAI Conference on Artificial Intelligence*, 9153–9161. <https://doi.org/10.1609/aaai.v37i8.26098>
24. Deihim, A., Alonso, E., Apostolopoulou, D. (2023). STTRE: A Spatio-Temporal Transformer with Relative Embeddings for Multivariate Time Series Forecasting. *Neural Networks*, 168, 549–559.
25. Shi, J., Ma, Q., Ma, H., Li, L. (2024). Scaling Law for Time Series Forecasting. *Advances in Neural Information Processing Systems (NeurIPS)*.
26. Dehghani, M., Djolonga, J., Mustafa, B., et al. (2023). Scaling Vision Transformers to 22 Billion Parameters. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 7480–7512.
27. Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference*

- of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (Volume 1: Long Papers), 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
28. Loshchilov, I., Hutter, F. (2019). Decoupled Weight Decay Regularization. International Conference on Learning Representations (ICLR). arXiv:1711.05101
29. OpenAI. (2023). GPT-4 Technical Report. <https://cdn.openai.com/papers/gpt-4.pdf>
30. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2980–2988. <https://doi.org/10.1109/ICCV.2017.324>
31. Sun, C., Li, H., Song, M., Cai, D., Hong, S. (2023). Curricular and Cyclical Loss for Time Series Learning Strategy. CoRR, abs/2312.15853.
32. Sun, C., Li, H., Song, M., Cai, D., Zhang, B., Hong, S. (2023). Continuous Diagnosis and Prognosis by Controlling the Update Process of Deep Neural Networks. Patterns, 4(2), 100687. <https://doi.org/10.1016/j.patter.2023.100687>
33. Jing, J., Lin, Z., Yang, C., et al. (2024). HMS - Harmful Brain Activity Classification. Kaggle Competition. <https://kaggle.com/competitions/hms-harmful-brain-activity-classification>
34. Westover, M. B., Shafi, M. M., Ching, S., et al. (2013). Real-time segmentation of burst suppression patterns in critical care EEG monitoring. Journal of Neuroscience Methods, 219(1), 131–141. <https://doi.org/10.1016/j.jneumeth.2013.07.003>
35. Perslev, M., Darkner, S., Kempfner, L. et al. (2021). U-Sleep: Resilient High-Frequency Sleep Staging. npj Digit. Med., 4, 72.
36. Chen, Y., Ren, K., Song, K., Wang, Y., Wang, Y., Li, D., Qiu, L. (2024). EEGFormer: Towards Transferable and Interpretable Large-Scale EEG Foundation Model. arXiv, abs/2401.10278.
37. Wang, C., Subramaniam, V., Yaari, A. U., Kreiman, G., Katz, B., Cases, I., Barbu, A. (2023). BrainBERT: Self-Supervised Representation Learning for Intracranial Recordings. arXiv preprint arXiv:2302.14367.
38. Yang, C., Xiao, C., Westover, M. B., et al. (2023). Self-Supervised Electroencephalogram Representation Learning for Automatic Sleep Staging: Model Development and Evaluation Study. JMIR AI, 2(1), e46769. <https://doi.org/10.2196/46769>
39. Li, R., Liu, L., Lu, B. (2021). Discrimination of Decision Confidence Levels from EEG Signals. 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER), 946–949. <https://doi.org/10.1109/NER49283.2021.9441086>
40. Kahneman, D., Rosenfield, A., Gandhi, L., Blaser, T. (2016). Noise. Harvard Business Review, 38–46.
41. Younes, M., Raneri, J., Hanly, P. (2016). Staging sleep in polysomnograms: Analysis of inter-scorer variability. J Clin Sleep Med, 12, 885–894. <https://doi.org/10.5664/jcsm.5894>

