EEG-Bench: A Benchmark for EEG Foundation Models in Clinical Applications

Ard Kastrati
ETH Zurich
akastrati@ethz.ch

Josua Bürki
ETH Zurich
jbuerki@ethz.ch

Jonas Lauer
ETH Zurich
jlauer@ethz.ch

Cheng Xuan
ETH Zurich
cxuan@ethz.ch

Raffaele Iaquinto
ETH Zurich
iaquintr@ethz.ch

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

Abstract

We introduce a unified benchmarking framework focused on evaluating EEG-based foundation models in clinical applications. The benchmark spans 11 well-defined diagnostic tasks across 14 publicly available EEG datasets, including epilepsy, schizophrenia, Parkinson's disease, OCD, and mild traumatic brain injury. It features minimal preprocessing, standardized evaluation protocols, and enables side-by-side comparisons of classical baselines and modern foundation models. Our results show that while foundation models achieve strong performance in certain settings, simpler models often remain competitive, particularly under clinical distribution shifts. To facilitate reproducibility and adoption, we release all prepared data and code in an accessible and extensible format.

1 Introduction

The rise of foundation models has started to revolutionize healthcare, with large language models (LLMs) achieving impressive results in medical reasoning, and clinical summarization [1, 2]. While language remains the primary focus, recent work is increasingly exploring multimodal foundation models that extend to vision, audio, and biosignals. Electroencephalography (EEG) is an essential modality in clinical neurology, widely used to diagnose and monitor brain disorders such as epilepsy, sleep disorders, etct. It is non-invasive, cost-effective, and offers high temporal resolution, making it a valuable tool across diverse clinical environments. However, EEG signals are notoriously difficult to work with: they are noisy, low in spatial resolution, and highly variable across subjects and hardware setups. To address these issues, recent work has proposed EEG-specific foundation models adapted from advances in language [3], vision [4], and audio processing [5]. Models such as BENDR [6], Neuro-GPT [7], and LaBraM [8] leverage large-scale public EEG datasets and self-supervised learning objectives to build general-purpose neural representations that transfer across subjects, tasks, and datasets. For evaluation of EEG foundation models, they are often fine-tuned on tasks from the brain-computer interface domain, as well as on clinical datasets like TUEG Epilepsy [9], TUAB Abnormal EEG detection [10], and sleep staging. However, EEG is also used to support diagnosis of many other conditions, including mild traumatic brain injury (mTBI), Parkinson's disease (PD), schizophrenia, and obsessive-compulsive disorder (OCD), which remain underrepresented in current evaluation efforts.

To address this, we present a unified benchmarking framework to evaluate EEG models across a wide range of clinical tasks and datasets. It includes 14 publicly available datasets spanning diverse clinical conditions, subject demographics, recording paradigms (e.g., sleep, rest), and hardware setups (e.g., different EEG caps and sampling rates), enabling evaluation under realistic and heterogeneous

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

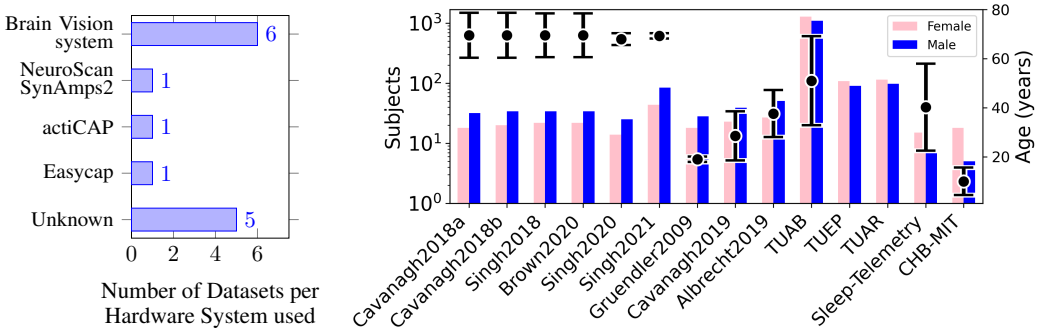

Figure 1: Dataset diversity across three dimensions: (Left) Distribution of EEG hardware systems used across datasets, each with potentially different electrode layouts, channel counts, and sampling rates. (Right) Gender and age distribution of subjects, covering a broad range from infants (1 year old) to elderly adults (up to 80 years), reflecting the inclusion of both pediatric and geriatric populations. The age bars show mean and standard deviation.

conditions (Figure 1, Table 1). From this data, we define 11 diagnostic tasks of varying difficulty (Table 2). We then perform the first comprehensive comparison of EEG foundation models and classical baselines in the clinical setting. Our benchmark provides a consistent, extensible platform for evaluating generalization and highlights the current strengths and limitations of general-purpose EEG models. All datasets and code are released in a reproducible, plug-and-play format to accelerate research in clinical EEG modeling and support the development of robust and generalizable neural decoding systems.

2 Benchmark Framework

To systematically evaluate the generalization capabilities of EEG decoding models, we introduce a benchmarking framework that integrates diverse datasets, defines a standardized set of decoding tasks and evaluation metrics, and supports evaluation across model types and domains.

Datasets. Our benchmark integrates 14 publicly available EEG datasets spanning multiple clinical paradigms, selected to reflect a wide range of real-world conditions. This includes heterogeneity in subject demographics, experimental protocols, clinical conditions, and recording setups and lengths. In this work, we focus on datasets (Table 1) that are primarily collected in medical contexts for diagnostic and monitoring purposes.

Benchmarking Tasks To assess EEG foundation models in realistic healthcare settings, we define a suite of 11 clinical tasks (Table 2). All tasks are evaluated in a strict *cross-subject* setting to ensure generalization beyond individual patients. The tasks cover both diagnostic classification and event detection. Diagnostic tasks involve full-length EEG recordings and include binary problems such as normal vs. abnormal or epileptic vs. non-epileptic, as well as disease-specific classifications for Parkinson's disease (PD), obsessive–compulsive disorder (OCD), schizophrenia, and mild traumatic brain injury (mTBI). As we have multiple datasets for PD, we additionally test the models on a dataset they were not trained on (held-out), for *cross-dataset* evaluation. Event-based tasks operate on shorter EEG segments and include seizure detection, artifact detection, and sleep stage classification. These tasks reflect the diversity of real-world clinical EEG use, with recordings ranging from minutes to hours, often involving uncontrolled cognitive states and substantial inter-subject variability. Together, they provide a challenging benchmark for evaluating whether foundation models can capture clinically meaningful neural representations.

Metrics We report *balanced accuracy* and *weighted F1 score* as our primary metrics, computed on the fixed test splits. These metrics are chosen for their robustness to class imbalance and wide adoption in EEG decoding literature, respectively.

Our benchmarking framework supports both standard machine learning models that rely on hand-crafted features and modern foundation models trained on large-scale EEG data.

Dataset	Task	Experimental Paradigm	Channels	Total Length	Subjects	Recordings
Cavanagh2018a [11]	PD	Oddball: 440Hz, 600Hz, novel; x100 trials, x2 blocks; 200ms; 2m	64	21 h	50	77
Cavanagh2018b [11]	PD		64	4 h	56	56
Cavanagh2019 [12]	mTBI		64	57 h	62	62
Albrecht2019 [13]	Schizo-phrenia	Simon task: > x80 trials, x4 blocks; 850ms	64	51 h	78	78
Singh2018 [14]	PD	Simon task: > x20 stimulus, x4 blocks; 750ms-1.5s	64	108 h	56	56
Brown2020 [15]	PD	Reward learning: No Choice A_{nc} B_{nc} 90% / 10%; C_{nc} D_{nc} 70% / 30%; Free Choice A_{fc} B_{fc} 90% / 10%; C_{fc} D_{fc} 70% / 30%; x20 stimulus, x3-x5 blocks; 4s	64	31 h	56	84
Gruendler2009 [16]	OCD	Reward learning: ちと せみ のま ちせ との みま; A (80%) B (20%) C (70%) D (30%) E (60%) F (40%); x60 trials, x6 blocks; 4s; testing phase; 4s	64	22 h	46	46
Singh2020 [17]	PD	0.5s; 1-2s; 3s; > x2 blocks, x30-x50 trials	64	7 h	39	39
Singh2021 [18]	PD	Instruction text: time-interval (3 or 7s); spacebar; 1s; x80 trials, x4 blocks; 8-20s	64	58 h	120	129
TUAB [10]	Abnormal	*Data collected from patients in hospital beds during monitoring*	19	47.5 d	2,383	2,993
TUEP [9]	Epilepsy		19	26.3 d	200	2,041
TUAR [19]	Artifact		19	4.2 d	213	310
CHB-MIT [20]	Seizure		18	41 d	23	686
Sleep-Telemetry [21]	Sleep Stages	temazepam; 1night; placebo; 1night	3	15.8 d	22	44

Table 1: Clinical Datasets. PD stands for Parkinson's Disease, mTBI for mild Traumatic Brain Injury and OCD for Obsessive-Compulsive Disorder. stands for eye open, sleeping, 440Hz 600Hz novel oddball stimulus, modified Simon task with the shapes and colors, pedal press, space bar press, hospital patient monitoring.

Standard ML Models As classical baselines, we use Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) applied to handcrafted features. All signals are resampled to 200 Hz, restricted to the subset of channels common across datasets, and truncated to a consistent length. Features are extracted with the Brainfeatures toolbox, which provides a comprehensive set of time- and frequency-domain measures. These include spectral power across standard frequency bands (delta–gamma), statistical moments, and complexity metrics such as entropy and fractal dimension. Transformations such as the continuous and discrete wavelet transforms (CWT, DWT) and the discrete Fourier transform (DFT) are applied to capture rich temporal–spectral dynamics. Brainfeatures also supports epoching and aggregation, enabling it to handle recordings of arbitrary length and generate consistent session-level representations, which is critical for heterogeneous clinical EEG.

Task	Class	#Samples
Abnormal	Abnormal	1,472
	Normal	1,521
Epilepsy	Epilepsy	1,651
	No Epilepsy	390
PD (All)	Parkinson's	266
	Control	157
PD (Held-Out)	Parkinson's	194
	Control	115
OCD	OCD	22
	Control	24
mTBI	mTBI	104
	Control	73
Schizophrenia	Schizophrenia	45
	Control	31

Task	Class	#Samples
Binary Artifact	No Event	216,466
	Artifact	141,102
Multiclass Artifact	No Event	216,466
	Eye Movement	34,480
	Muscle Artifact	60,763
	Electrode Artifact	41,652
	Chewing	3,964
	Shivers	243
Sleep Stages	Unknown Phase	82,734
	Wake Phase	132,746
	N1 Phase	109,590
	N2 Phase	595,530
	N3 & N4 Phase	192,450
	REM Phase	250,470
Seizure	No Seizure	3,515,547
	Seizure	11,525

Table 2: Overview of the benchmarking tasks, showing the classes each task distinguishes. #Samples denotes the number of samples of each class. Here, PD stands for Parkinson's Disease, OCD for Obsessive-Compulsive Disorder and mTBI for mild Traumatic Brain Injury.

Foundation Models We also evaluate foundation models that learn EEG representations directly from raw signals: BENDR [6], Neuro-GPT [7], and LaBraM [8]. For all, signals are bandpass filtered (0.1–75 Hz), notch filtered to suppress line noise, and resampled to 200 Hz. BENDR and Neuro-GPT require a fixed channel set; for each dataset, we hence retain the pretrained channels, zero-pad missing ones, and map datasets with incompatible layouts where necessary. We prepend every released model to a linear classification head and fine-tune the entire model. For Neuro-GPT, following the authors' findings, we fine-tune only the encoder. LaBraM is more flexible with respect to channel configurations and can thus accommodate heterogeneous datasets. We use the publicly released LaBraM-Base and fine-tune it end-to-end. Since clinical recordings often span minutes to hours, exceeding the input limits of these models (4–60 s), we segment each recording into non-overlapping chunks, encode them separately, and average embeddings before classification. This allows scalable inference on long-duration EEG while retaining global context. More details on how we used these models can be found in appendix B.

3 Experiments

We evaluate the generalization performance of both classical and foundation models across the EEG decoding tasks defined in our benchmark. Each experiment was repeated five times with different random seeds, requiring 270 hours on an A100 GPU and 16 AMD EPYC 7742 CPUs in total. We report the mean and standard deviation of the balanced accuracy scores in Table 3. Corresponding weighted F1 scores are provided in Table 5 in the appendix for completeness.

We can see in Table 3 that LaBraM frequently achieved the highest or competitive performance across a range of tasks. For example, it reached a balanced accuracy of 0.838 in abnormal EEG detection — substantially outperforming all other models — and delivered strong results on Obsessive-Compulsive Disorder (OCD) classification (0.740). These results demonstrate LaBraM's ability to model long, noisy, and heterogeneous EEG recordings, even under varied clinical conditions. Interestingly, epilepsy detection presented an exception. Despite LaBraM's strong performance on many other clinical tasks, it underperformed here with only 0.565 balanced accuracy, while BENDR and Neuro-GPT achieved much stronger results (0.740 and 0.734, respectively). Given the highly imbalanced nature of the epilepsy dataset, this suggests that LaBraM may be more sensitive to label imbalance or more prone to overfitting in such contexts. In the mild traumatic brain injury (mTBI) task, a simple classical method — LDA — surprisingly outperformed all foundation models, reaching a balanced accuracy of 0.813. This result highlights that in low-data regimes, classical models with strong inductive biases and low capacity can remain not only viable but also superior. All foundation models

Task	SVM	LDA	BENDR	Neuro-GPT	LaBraM
Abnormal	0.722	0.677	0.717 ± .003	0.696 ± .005	**0.838 ± .011**
Epilepsy	0.531	0.531	**0.740 ± .015**	0.734 ± .010	0.565 ± .017
PD (All)	0.648	0.658	0.529 ± .009	**0.687 ± .000**	0.656 ± .025
PD (Held-Out)	0.596	0.654	0.615 ± .038	**0.673 ± .000**	0.673 ± .038
OCD	0.633	0.717	0.513 ± .051	0.703 ± .082	**0.740 ± .044**
mTBI	0.626	**0.813**	0.640 ± .093	0.646 ± .000	0.740 ± .173
Schizophrenia	**0.679**	0.547	0.471 ± .055	0.545 ± .042	0.543 ± .045
Binary Artifact	0.745	0.705	0.535 ± .003	0.711 ± .004	**0.756 ± .007**
Multiclass Artifact	**0.437**	0.325	0.192 ± .002	0.226 ± .006	0.430 ± .015
Sleep Stages	0.652	**0.671**	0.169 ± .001	0.166 ± .003	0.192 ± .001
Seizure	0.572	0.529	0.501 ± .001	0.500 ± .000	**0.588 ± .011**

Table 3: Balanced Accuracy scores achieved by all models across evaluated tasks. Features for SVM or LDA were extracted using the Brainfeatures toolbox. For foundation models, we always report "mean ± standard-deviation" among the five runs.

struggled on this task, likely due to their high capacity and lack of regularization in small-sample settings. In schizophrenia classification, SVM was the best-performing model (0.679), with all foundation models, including Neuro-GPT (0.545) and LaBraM (0.543), falling behind. None of the deep models learned meaningful patterns in this task — likely due to the subtle and distributed nature of schizophrenia-related EEG markers, which can be difficult to capture without tailored inductive structure or more specialized training data. Throughout the multi-label tasks, we observe that SVM, LDA and LaBraM by and large yielded the highest scores, with at least one of the baseline models never being more than 2% worse than LaBraM. However, for sleep stages, the baselines yielded stable results around 0.66 balanced accuracy, while the foundation models severely struggled, with LaBraM – the best among them – achieving a score of merely 0.192. The inability of BENDR and Neuro-GPT to learn anything beyond random guessing for the sleep stages and seizure tasks might suggest that these models are not capable of using previously unseen channels in a meaningful way, as the datasets included in each task have no channels in common with either model. Despite not having been trained on the datasets' channels, LaBraM was able to advance to at least a few percentage points above random guesses, indicating a greater flexibility with respect to channels.

4 Conclusion

As EEG foundation models continue to emerge, robust and clinically meaningful evaluation becomes increasingly important. In this work, we introduce a unified benchmarking framework for EEG decoding in healthcare, with a focus on diagnostic and event-based clinical tasks. By standardizing access to 14 publicly available datasets and defining 11 diverse classification tasks, our benchmark provides a transparent and extensible platform for evaluating models in realistic clinical settings.

Our contributions are threefold: *(i) Standardized and Extensible Framework*: A unified platform that parses clinical EEG datasets into a common format, applies minimal preprocessing, and supports both classical and modern model evaluation. *(ii) Well-defined Clinical Tasks:* A curated suite of 11 diagnostic and event-based tasks spanning a range of neurological and psychiatric conditions. *(iii) Comprehensive Performance Evaluation:* Side-by-side comparisons of classical baselines and EEG foundation models to assess generalization across subjects and clinical conditions.

Limitations and Future Work. While our primary focus has been on building and validating the benchmarking framework, the current model suite includes only a few standard ML pipelines and foundation models. Many specialized models have been developed for individual tasks (e.g., seizure detection, sleep staging, schizophrenia diagnosis), and integrating these will provide a more complete view of model capabilities and limitations. We plan to continuously expand the benchmark to include more datasets, tasks, and models and encourage the community to contribute evaluations of emerging EEG models under this unified setup. By providing a rigorous, transparent, and extensible benchmark, we hope to foster more reliable progress in EEG decoding and support the development of foundation models that truly generalize across the diverse landscape of EEG data in clinical setups.

References

[1] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[2] Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025.

[3] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[4] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundational models defining a new era in vision: A survey and outlook, 2023.

[5] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen, 2023.

[6] Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. *Frontiers in Human Neuroscience*, 15, 2021.

[7] Wenhui Cui, Woojae Jeong, Philipp Thölke, Takfarinas Medani, Karim Jerbi, Anand A. Joshi, and Richard M. Leahy. Neuro-gpt: Towards a foundation model for eeg, 2024.

[8] Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci, 2024.

[9] L. Veloso, J. McHugh, Eva von Weltin, Sebas Lopez, I. Obeid, and Joseph Picone. Big data resources for eegs: Enabling deep learning research. pages 1–3, 12 2017.

[10] Iyad Obeid and Joseph Picone. The temple university hospital eeg data corpus. *Frontiers in Neuroscience*, 10, 2016.

[11] James F. Cavanagh, Praveen Kumar, Andrea A. Mueller, Sarah Pirio Richardson, and Abdullah Mueen. Diminished eeg habituation to novel events effectively classifies parkinson's patients. *Clinical Neurophysiology*, 129(2):409–418, 2018.

[12] J. F. Cavanagh, J. K. Wilson, R. E. Rieger, D. Gill, J. M. Broadway, J. H. Story Remer, V. Fratzke, A. R. Mayer, and D. K. Quinn. Erps predict symptomatic distress and recovery in sub-acute mild traumatic brain injury. *Neuropsychologia*, 132:107–125, September 2019. Epub 2019 Jun 19.

[13] Matthew A. Albrecht, James A. Waltz, James F. Cavanagh, Michael J. Frank, and James M. Gold. Increased conflict-induced slowing, but no differences in conflict-induced positive or negative prediction error learning in patients with schizophrenia. *Neuropsychologia*, 123:131–140, 2019. Cognitive Effort.

[14] A. Singh, S. P. Richardson, N. Narayanan, and J. F. Cavanagh. Mid-frontal theta activity is diminished during cognitive control in parkinson's disease. *Neuropsychologia*, 117:113–122, August 2018. Epub 2018 May 23.

[15] D. R. Brown, S. P. Richardson, and J. F. Cavanagh. An eeg marker of reward processing is diminished in parkinson's disease. *Brain Research*, 1727, January 2020. Epub 2019 Nov 5.

[16] T. O. Gründler, J. F. Cavanagh, C. M. Figueroa, M. J. Frank, and J. J. Allen. Task-related dissociation in ern amplitude as a function of obsessive-compulsive symptoms. *Neuropsychologia*, 47(8-9):1978–1987, July 2009. Epub 2009 Mar 17.

[17] Arun Singh, Rachel C. Cole, Arturo I. Espinoza, Darin Brown, James F. Cavanagh, and Nandakumar S. Narayanan. Frontal theta and beta oscillations during lower-limb movement in parkinson's disease. *Clinical Neurophysiology*, 131(3):694–702, 2020.

[18] A. Singh, R. C. Cole, A. I. Espinoza, A. Evans, S. Cao, J. F. Cavanagh, and N. S. Narayanan. Timing variability and midfrontal $\sim$4 hz rhythms correlate with cognition in parkinson's disease. *NPJ Parkinson's Disease*, 7(1):14, February 2021. Published 2021 Feb 15.

[19] Ahmed Hamid, Katherine Gagliano, Safwanur Rahman, Nikita Tulin, Vincent Tchiong, Iyad Obeid, and Joseph Picone. The temple university artifact corpus: An annotated corpus of eeg artifacts. In *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–4. IEEE, 2020.

[20] Ali Hossam Shoeb. *Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment*. PhD thesis, Massachusetts Institute of Technology, 2009.

[21] M. S. Mourtazaev, B. Kemp, A. H. Zwinderman, and H. A. Kamphuisen. Age and gender affect different characteristics of slow waves in the sleep eeg. *Sleep*, 18:557–564, 1995.

[22] Silvia López de Diego. Automated interpretation of abnormal adult electroencephalograms. Master's thesis, Temple University, 2017.

[23] World Health Organization. Epilepsy: A public health imperative, 2019. Geneva.

A EEG-Bench Interface

We have made the code of our benchmarking tool publicly available at `https://github.com/ETH-DISCO/EEG-Bench`, licensed under the GNU GPL v3.0 license or later.

Running the benchmark

To run the benchmark, create the conda environment `eeg_bench` via

```
conda env create -f environment.yml
conda activate eeg_bench
```

Then, configure the paths to your local storage in `eeg_bench/config.json` and run the benchmark via

```
python benchmark_console.py --all
```

Dataset Accessibility and Distribution

A core design principle of EEG-Bench is ease of use: To foster widespread adoption and reproducibility, we aim to abstract away the often cumbersome process of data acquisition. Thus, we implemented automated download mechanisms for public datasets using custom scripts integrated into our codebase. This ensures that users can seamlessly access all of the 14 datasets with minimal effort and (most of the time) no need for manual intervention.

Notably, datasets hosted by the Temple University Hospital (TUAB, TUEP, TUAR) require acceptance of a data use agreement (DUA) via the NEDC portal; once accepted, they can be automatically retrieved using our tools.

The table below summarizes dataset accessibility across all 14 benchmark datasets.

Dataset Name	Access Type	Notes / Requirements
TUAB	Automatic	DUA via NEDC; automatic download afterward
TUEP	Automatic	DUA via NEDC; automatic download afterward
TUAR	Automatic	DUA via NEDC; automatic download afterward
CHB-MIT	Automatic	
Sleep-Telemetry	Automatic	
Cavanagh2018a	Automatic	
Cavanagh2018b	Automatic	
Cavanagh2019	Automatic	
Singh2018	Automatic	
Singh2020	Automatic	
Singh2021	Automatic	
Brown2020	Automatic	
Gruendler2009	Automatic	
Albrecht2019	Automatic	

Table 4: Dataset Accessibility Overview

Adding Your Own Dataset

Our benchmarking code allows users to easily add their own datasets to the benchmark. Base classes are available for both, clinical and BCI datasets.

To add your dataset, follow these steps:

1. Place your dataset class in `eeg_bench/datasets/bci/` or `eeg_bench/datasets/clinical/`.
2. Inherit from `BaseBCIDataset` or `BaseClinicalDataset`.

3. Implement the following methods:
   (a) `_download`: Either download the dataset automatically or provide instructions for the user to do so manually. Pay attention that, if possible, `_download` does not re-download the dataset if it already exists locally.
   (b) `load_data`: This method should populate the following attributes:
      - `self.data`: `np.ndarray` or `List[BaseRaw]` with shape $(n_{\text{samples}}, n_{\text{channels}}, n_{\text{sample length}})$
      - `self.labels`: `np.ndarray` or `List[str]` with shape $(n_{\text{samples}},)$, or $(n_{\text{samples}}, n_{\text{multi_labels}})$ for multi-label datasets
      - `self.meta`: A dictionary that must contain at least `name`, `sampling_frequency` and `channel_names`
   (c) If your dataset contains classes not yet defined in the enums `enums.BCIClasses` or `enums.ClinicalClasses`, please add them accordingly.
   (d) For multi-label datasets, you currently also have to add your dataset name to the
   ```
   elif dataset_name in [<MULTILABEL_DATASET_NAMES>]:
   ```
   clause in `eeg_bench/models/clinical/brainfeatures/feature_extraction_2.py:_prepare_data_cached()`.
   (e) To speed up further runs of the `load_data` function, implement caching as in the existing dataset classes.
   (f) All EEG signals should be standardized to the microvolt (μV) scale. To reduce memory usage and computational overhead, signals with a sampling rate greater than 250 Hz are typically resampled to 250 Hz.

Adding Your Own Task

Tasks are the central organizing principle of the benchmark, encapsulating paradigms, datasets, prediction classes, subject splits (i.e., training and test sets), and evaluation metrics. Each task class implements a `get_data()` method that returns training or testing data, along with the corresponding labels and metadata. These predefined splits ensure evaluation consistency and facilitate reproducibility. The tasks are divided into **Clinical** and **BCI** categories.

Each task defines:

- The datasets to be used
- Training and testing subject splits
- Target classes
- Evaluation metrics

To add your own task:

- For BCI tasks, add your class to `tasks/bci/` and inherit from `AbstractBCITask`
- For Clinical tasks, add your class to `tasks/clinical/` and inherit from `AbstractClinicalTask`

You must implement the `get_data()` method to return training or testing splits along with data, labels, and metadata.

For multi-label tasks, you must also add its name to the `get_multilabel_tasks()` method in `eeg_bench/utils/utils.py`. Additionally, if you have special channel requirements, you might also want to add an

```
elif task_name == <YOUR_TASK_NAME>:
    t_channels = <YOUR_CHANNEL_LIST>
```

clause to `_prepare_data_cached()` in `eeg_bench/models/clinical/brainfeatures/feature_extraction_2.py`.

Add Your Own Model

To integrate a new model into the benchmark, implement the `AbstractModel` interface and place your implementation in the appropriate directory:

- `models/bci/` for Motor Imagery (BCI) models
- `models/clinical/` for Clinical models

Required Methods

Your model must implement the following methods:

```
def fit(self, X: List[np.ndarray | List[BaseRaw]],
              y: List[np.ndarray | List[str]],
              meta: List[Dict]) -> None:
    # Each list entry corresponds to one dataset
    pass

def predict(self, X: List[np.ndarray | List[BaseRaw]],
                  meta: List[Dict]) -> np.ndarray:
    # Predict on each dataset separately, return concatenated predictions
    pass
```

Running Your Model

To run your model, register it in `benchmark_console.py` and execute the following command:

```
python benchmark_console.py --model mymodel --task <YOUR_DESIRED_TASK>
```

B Models

In this section, we provide a more detailed description of all the baselines and the foundation models used for this paper.

B.1 Standard ML Baselines

To establish strong baselines for the benchmarking tasks, we employed classical machine learning pipelines that are well-established in EEG research. Specifically, we used Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM), which operate on handcrafted features derived from the EEG signal. These models are known for their robustness in low-data regimes and their interpretability, making them a natural starting point for benchmarking. To ensure uniformity across datasets, for standard ML baselines, all EEG signals are resampled to 200 Hz. Since datasets differ in both channel layout and recording duration, for standard ML models, we restrict our analysis to the subset of channels common across datasets and truncate signals to a consistent length. Following this, we apply the feature extraction techniques described below.

Feature Extraction in Clinical Tasks: For our clinical classification tasks, which include long-duration recordings and full-recording-level predictions, we utilize the Brainfeatures toolbox[1]. This open-source tool provides a comprehensive set of EEG features derived from both time and frequency domains. It applies a variety of transformations—including the continuous wavelet transform (CWT), discrete wavelet transform (DWT), and discrete Fourier transform (DFT)—to extract rich representations of EEG dynamics. Extracted features include spectral power across standard frequency bands (delta: 0.5–4 Hz, theta: 4–8 Hz, alpha: 8–13 Hz, beta: 13–30 Hz, and gamma: 30–100 Hz), statistical moments (mean, variance, higher-order moments), as well as complexity metrics such as entropy and fractal dimension. Crucially, Brainfeatures includes robust mechanisms for epoching and aggregating features over time, enabling it to process EEG recordings of arbitrary length and generate consistent, session-level feature representations. This makes it particularly suitable for clinical applications where recording conditions are more heterogeneous.

[1] Code available at `https://github.com/TNTLFreiburg/brainfeatures`

B.2 Large Pretrained Models

In addition to classical pipelines, our benchmark incorporates large pretrained deep learning models designed to learn generalizable EEG representations directly from raw signals with minimal preprocessing. We evaluate three recent architectures: BENDR, Neuro-GPT, and LaBraM.

BENDR BENDR [6] requires a fixed set of input channels. Therefore, we select only the subset of channels used during pretraining. If any of these channels are missing in a given dataset, we zero-pad the corresponding input. An exception to this rule comes into effect when *all* channels are missing, as is the case in the Sleep-Telemetry and CHB-MIT datasets. In this case, we assign the dataset channels to arbitrary BENDR input channels. We apply a bandpass filter from 0.1 to 75.0 Hz and a notch filter to suppress power line artifacts. All signals are resampled to 200 Hz to match the model's expected input rate. We use the publicly released pretrained BENDR encoder and append a linear classification head. The entire model, including the encoder, is fine-tuned on each task.

Neuro-GPT Similar to BENDR, Neuro-GPT [7] requires a fixed channel layout, so we apply the same preprocessing steps, including channel selection, bandpass and notch filtering, and resampling to 200 Hz. We utilize the pretrained model released by the authors and follow their recommended approach of fine-tuning the encoder only, as it was shown to yield the best performance in their experiments.

LaBraM LaBraM [8] is more flexible with respect to channel configurations, making it more suitable for heterogeneous EEG datasets. As with the other models, we apply a 0.1–75.0 Hz bandpass filter, a notch filter, and resample all signals to 200 Hz. We use the LaBraM-Base variant, which is the only version publicly released, and fine-tune it end-to-end on our tasks.

Handling Long Clinical Recordings Clinical datasets often contain continuous recordings spanning several minutes or hours, exceeding the input length limits of these models (ranging from 4 to 60 seconds, depending on the architecture). To address this, we divide each sample exceeding these length limits into N non-overlapping chunks of fixed duration. Each chunk is passed through the encoder, and the resulting embeddings are averaged across all chunks before being fed into the final classification layer. This strategy allows efficient and scalable inference over long-duration EEG data while preserving global context through embedding aggregation.

C Weighted F1-Score Results and Task Details

In the following table, we report the results of our experiments from Section 3 also using weighted F1 score.

Task	SVM	LDA	BENDR	Neuro-GPT	LaBraM
Abnormal	0.720	0.680	0.722 ± .003	0.699 ± .005	**0.842 ± .012**
Epilepsy	0.613	0.593	**0.709 ± .029**	0.697 ± .016	0.647 ± .018
PD (All)	0.670	0.682	0.560 ± .016	**0.724 ± .000**	0.692 ± .021
PD (Held-Out)	0.662	0.707	0.683 ± .043	0.729 ± .000	**0.742 ± .037**
OCD	0.636	0.723	0.372 ± .127	0.681 ± .089	**0.743 ± .042**
mTBI	0.580	**0.793**	0.704 ± .093	0.683 ± .000	0.776 ± .152
Schizophrenia	**0.681**	0.533	0.421 ± .084	0.544 ± .041	0.463 ± .128
Binary Artifact	**0.761**	0.728	0.554 ± .003	0.723 ± .006	0.752 ± .009
Multiclass Artifact	**0.683**	0.643	0.430 ± .002	0.427 ± .012	0.624 ± .017
Sleep Stages	0.700	**0.732**	0.245 ± .001	0.231 ± .010	0.264 ± .002
Seizure	0.974	0.987	0.994 ± .000	**0.995 ± .000**	0.986 ± .003

Table 5: Weighted F1-scores achieved by all models across evaluated tasks. Features for SVM or LDA were extracted using the Brainfeatures toolbox.

Additionally, the following table specifies the datasets included in the evaluation set of each task.

Dataset	Abnormal	Epilepsy	PD (All)	PD (Held-Out)	OCD	mTBI	Schizophrenia	Binary Artifact	Multiclass Artifact	Sleep Stages	Seizure
TUAB	✓										
TUEP		✓									
Cavanagh2018a			✓								
Cavanagh2018b			✓								
Singh2018			✓								
Brown2020			✓								
Singh2020			✓	✓							
Singh2021			✓								
Gruendler2009					✓						
Cavanagh2019						✓					
Albrecht2019							✓				
TUAR								✓	✓		
Sleep-Telemetry										✓	
CHB-MIT											✓

Table 6: Overview of clinical dataset inclusion. PD=Parkinson's disease, OCD=Obsessive-Compulsive Disorder, mTBI=mild Traumatic Brain Injury.

Lastly, Table 7 gives more detailed information about each of the tasks that we defined.

D Datasets

Clinical datasets reflect neural conditions that arise naturally or as a result of neurological or psychiatric disorders. These datasets are valuable for automating the detection of pathological events and conditions such as seizures, epilepsy, Parkinson's disease, and schizophrenia, as well as non-pathological states like sleep stages. They are often collected in hospital settings and, as a result, they

Task		Class	Train Samples	Test Samples	Sample Length
Abnormal		Abnormal	1,346	126	22.9 m
		Normal	1,371	150	
Epilepsy		Epilepsy	1,384	267	18.4 m
		No Epilepsy	268	122	
Parkinson's Disease (PD)	(All)	Parkinson's	168	98	24.8 m
		Control	102	55	
	(Held-Out)	Parkinson's	168	26	24.1 m
		Control	102	13	
Obsessive-Compulsive Disorder (OCD)		OCD	17	5	22.6 m
		Control	18	6	
mild Traumatic Brain Injury (mTBI)		mTBI	85	19	16.5 m
		Control	64	9	
Schizophrenia		Schizophrenia	38	7	27.0 m
		Control	25	6	
Binary Artifact		No Event	160,879	55,587	16 s
		Artifact	107,457	33,645	
Multiclass Artifact		No Event	160,879	55,587	16 s
		Eye Movement	27,196	7,284	
		Muscle Artifact	44,451	16,312	
		Electrode Artifact	32,104	9,548	
		Chewing	3,494	470	
		Shivers	212	31	
Sleep Stages		Unknown Phase	65,794	16,940	16 s
		Wake Phase	101,546	31,200	
		N1 Phase	84,300	25,290	
		N2 Phase	443,070	152,460	
		N3 & N4 Phase	150,900	41,550	
		REM Phase	191,190	59,280	
Seizure		No Seizure	2,548,702	966,845	16 s
		Seizure	8,514	3,011	

Table 7: Detailed overview of the benchmarking tasks, displaying the number of samples per class and train/test subset, as well as the average sample length.

tend to exhibit higher levels of noise and variability than, for example, BCI datasets recorded in a laboratory environment. While this introduces challenges for the model to learn in spite of this noise, it can also improve the robustness and generalizability of models trained on such data for real-world applications. Below, we provide an overview of the clinical datasets included in our benchmark.

D.1 Cavanagh2018a

The Cavanagh2018a dataset originates from an EEG-based clinical study investigating neural habituation to novel auditory stimuli in Parkinson's disease (PD) patients and matched controls [11]. The primary goal was to evaluate whether EEG responses to novelty could serve as a biomarker for cognitive dysfunction in PD.

A total of 53 participants were involved: 25 individuals diagnosed with Parkinson's disease (mean age: 69.68 ± 8.73 years; 16 male, 9 female) and 28 healthy age- and sex-matched control subjects. Each PD participant completed two sessions: one ON dopaminergic medication and one OFF (after 15 hours of medication withdrawal). Control participants completed a single session. All sessions were conducted at 9 AM to reduce circadian variability.

Participants underwent a three-stimulus auditory oddball task while EEG was recorded from 64 Ag/AgCl scalp electrodes placed according to the international 10/20 system. The task consisted of two blocks of 100 trials each, with three types of stimuli: frequent standard tones (440 Hz, 70% of trials), infrequent target tones (660 Hz, 15%), and novel sounds (15%) taken from naturalistic audio recordings. Each stimulus lasted 200 ms, and the average task duration per session was approximately 12 minutes.

The participants were instructed to silently count the target tones and ignore standards and novels. This passive paradigm was designed to isolate cognitive processing without motor response confounds.

Number of subjects	53 (25 PD, 28 Control)
Sessions per subject (average)	1.54
Average session length	16.4 min
EEG channels	64
Sampling rate	500 Hz
Total recordings	77
Hardware	Brain Vision system

Table 8: Summary of the Cavanagh2018a dataset

D.2 Cavanagh2018b

The Cavanagh2018b dataset contains resting-state EEG recordings from the same group of people used in the Cavanagh2018a study [11]. Participants included 28 individuals diagnosed with Parkinson's disease and 28 age- and sex-matched control participants. Each subject underwent a 2-minute resting-state EEG recording session with eyes open. These recordings were collected in a seated posture, prior to or following the auditory oddball task, under the same EEG setup (64-channel cap, 500 Hz sampling rate).

This dataset serves as a complementary baseline condition for evaluating spontaneous brain dynamics in Parkinson's disease. Though not central to the novelty task paradigm, resting-state data may be useful for future investigations into low-frequency oscillations or non-task-based classification approaches of Parkinson's disease.

Number of subjects	56 (28 PD, 28 Control)
Sessions per subject	1
Average session length	2 min
EEG channels	64
Sampling rate	500 Hz
Total recordings	56
Hardware	Brain Vision system

Table 9: Summary of the Cavanagh2018b dataset

D.3 Albrecht2019

The Albrecht2019 dataset originates from a study investigating reinforcement learning under cognitive conflict in individuals with schizophrenia (PSZ) and healthy controls [13]. The dataset includes both behavioral and EEG recordings collected during a modified Simon task, which introduces response conflict as an implicit cognitive cost during reinforcement learning.

A total of 78 participants took part in the study: 46 individuals with a DSM-IV diagnosis of schizophrenia or schizoaffective disorder, and 32 healthy controls. EEG data were recorded using a 64-channel Brain Vision system at a 1000 Hz sampling rate. Data were preprocessed and artifact-corrected using the EEGLAB pipeline, and ICA was applied for eye-blink removal. EEG epochs were extracted around stimulus and feedback events to capture conflict-evoked and prediction-error-related activity, particularly in the theta band (4–7 Hz).

Participants completed a reinforcement learning version of the Simon task. On each trial, a stimulus was associated with probabilistic reward or punishment outcomes, modulated by whether the trial

involved a congruent or conflict-inducing response. This design enabled the dissociation of positive and negative prediction error (PE) learning biases under cognitive effort. A subsequent transfer phase assessed stimulus preferences to infer learning outcomes.

Number of subjects	78 (46 PSZ, 32 Controls)
Sessions per subject	1
Average session length	40.3 min
EEG channels	64
Sampling rate	1000 Hz
Total recordings	76
Hardware	Brain Vision system

Table 10: Summary of the Albrecht2019 dataset

D.4 Singh2018

The Singh2018 dataset includes EEG recordings collected during a cognitive control task from 28 individuals with Parkinson's disease (PD) and 28 demographically matched healthy controls [14]. Each participant completed a modified Simon reaction-time task designed to elicit response conflict and error-related cognitive control processes. PD patients participated in two sessions (ON and OFF dopaminergic medication), spaced one week apart, while controls participated in a single session.

EEG was recorded from 64 scalp electrodes using a Brain Vision system at a sampling rate of 500 Hz.

Number of subjects	56 (28 PD, 28 Controls)
Sessions per subject	1
Average session length	118 min
EEG channels	64
Sampling rate	500 Hz
Total recordings	55
Hardware	Brain Vision system

Table 11: Summary of the Singh2018 dataset

D.5 Brown2020

The Brown2020 dataset comprises EEG recordings from a reinforcement learning task aimed at assessing reward processing in individuals with Parkinson's disease (PD) and healthy controls [15]. A total of 56 participants took part: 28 individuals diagnosed with PD and 28 age- and sex-matched control participants. Each PD participant completed two sessions (ON and OFF dopaminergic medication), spaced one week apart. Control participants completed a single session.

Participants performed a reinforcement learning task involving probabilistic feedback. On each trial, a pair of colored stimuli was presented, with each stimulus associated with a predefined probability of reward. Conditions were manipulated along two dimensions: difficulty (90/10% vs. 70/30% reward probability) and volition (free choice vs. instructed choice). The EEG was time-locked to the feedback screen, allowing for the measurement of reward-related event-related potentials (ERPs).

EEG data were recorded using a 64-channel Brain Vision system at a sampling rate of 500 Hz.

D.6 Gruendler2009

The dataset Gruendler2009 [16] originates from an EEG experiment that examined the relationship between obsessive–compulsive (OC) symptomatology and error-related brain activity. Participants were 46 undergraduate students selected based on their scores on the Obsessive–Compulsive Inventory-Revised (OCI-R), with groups categorized as high or low OC.

The experiment consisted of a flanker task to elicit ERNs from motor errors in a response conflict paradigm. Specifically, participants were shown a 5-letter string like "QQQQQ" or "QQOQQ".

Number of subjects	56 (28 PD, 28 Controls)
Sessions per subject	1.5
Average session length	22.1 min
EEG channels	64
Sampling rate	500 Hz
Total recordings	84
Hardware	Brain Vision system

Table 12: Summary of the Brown2020 dataset

Within a 1-second time-window, they then had to press a left or right button, depending on the letter in the middle of the string.

The EEG data were recorded using a 64-channel EEG + 2-channel EOG + 1-channel EKG setup. Participants were excluded for poor EEG quality, failure to meet learning criteria, or inconsistent OCI-R group classification. Demographic and psychometric data (including the Beck Depression Inventory) were collected to control for confounds.

Number of subjects	46
Sessions per subject	1
Average session length	28.7 min
EEG channels	64
Sampling rate	500 Hz
Total recordings	46
Hardware	NeuroScan SynAmps2

Table 13: Summary of the Gruendler2009 dataset

D.7 Cavanagh2019

The Cavanagh2019 dataset includes EEG recordings collected during a 3-stimulus auditory oddball paradigm in participants with mild traumatic brain injury (mTBI) and matched healthy controls [12]. A total of 85 participants took part: 38 sub-acute mTBI patients (tested within 2 weeks post-injury), 24 healthy controls, and 23 chronic TBI patients (mild to moderate severity). Sub-acute mTBI and control participants completed two or three EEG sessions – at 3–14 days and again after approximately 2 months – while chronic TBI participants completed a single session.

The task involved 260 trials: 70% standard tones (440 Hz), 15% target tones (660 Hz), and 15% novel naturalistic sounds. Stimuli were presented binaurally, and participants were instructed to count target tones while ignoring the others. EEG was recorded from 64 channels at a 500 Hz sampling rate.

Number of subjects	85 (38 sub-acute mTBI, 24 control, 23 chronic TBI)
Sessions per subject	2 (mTBI, controls), 1 (chronic TBI)
Average session length	22.1 min
EEG channels	64
Sampling rate	500 Hz
Total recordings	84
Hardware	Brain Vision system

Table 14: Summary of the Cavanagh2019 dataset

D.8 Singh2020

The Singh2020 dataset contains EEG recordings collected during a lower-limb pedaling task designed to assess motor control in individuals with Parkinson's disease (PD), with a particular focus on freezing of gait (FOG) symptoms [17]. A total of 39 participants were included: 13 PD patients with

FOG (PDFOG+), 13 PD patients without FOG (PDFOG-), and 13 demographically matched healthy controls.

Participants completed a lower-limb motor task in which they pedaled a stationary cycle in response to a visual "GO" cue, designed to minimize fall risk and reduce EEG artifacts from movement. Each subject completed at least two blocks of either 30 or 50 trials, with PDFOG+ participants performing fewer trials due to symptom severity. A tri-axial accelerometer mounted on the ankle measured pedaling kinematics, including mean speed and time to peak acceleration.

EEG was recorded using a 64-channel cap with a sampling rate of 500 Hz.

Number of subjects	39 (13 PDFOG+, 13 PDFOG-, 13 Controls)
Sessions per subject	1
Average session length	10.8 min
EEG channels	64
Sampling rate	500 Hz
Total recordings	39
Hardware	Easycap Brain Products

Table 15: Summary of the Singh2020 dataset

D.9 Singh2021

The Singh2021 dataset contains EEG recordings collected during an interval timing task designed to study cognitive control in individuals with Parkinson's disease (PD) [18]. A total of 130 participants were recruited: 89 PD patients and 41 demographically matched healthy controls. Of these, usable EEG data were available for 83 PD patients and 37 controls, after excluding sessions with insufficient data or poor signal quality. Most PD patients (n = 80) completed the task while ON medication, and a subset (n = 9) completed both ON and OFF dopaminergic medication sessions.

Participants performed a peak-interval timing task with intermixed 3-second and 7-second trials. They were instructed to press a key when they estimated the target interval had elapsed. Visual distractions were included to discourage counting. Each participant completed 80 trials (40 per interval type). Only trials with a minimum of 20 valid keypresses per interval condition were included in analyses.

EEG was recorded using a 64-channel actiCAP system at 500 Hz.

Number of subjects	120 (83 PD, 37 Controls)
Sessions per subject	1–2 (PD), 1 (Control)
Average session length	30.0 min
EEG channels	64
Sampling rate	500 Hz
Total recordings	129
Hardware	actiCAP Brain Products

Table 16: Summary of the Singh2021 dataset

D.10 TUAB

TUAB is the second-largest annotated subset of the TUEG corpus [10], collected at Temple University Hospital. It contains EEG recordings from 2,383 subjects that were classified as either "normal", if it fulfills certain characteristics, or as "abnormal", if it does not or if it contains patterns indicating pathological conditions [22].

This very basic type of classification can serve as an important first step in deciding whether to inspect a given EEG recording closer and search for more concrete conditions.

Table 17 gives an overview of the statistics of the TUAB dataset.

Number of subjects	2,383
Sessions per subject (average)	1.3
Average session length	23 min
EEG channels	19
Sampling rates	$\{250, 256, 512\}$ Hz
Number of abnormal recordings	1,472
Number of normal recordings	1,521

Table 17: Summary of the TUAB dataset

D.11 TUEP

Epilepsy is a neurological disorder affecting an estimated 50 million people world-wide [23]. In order to treat a patient with epilepsy, the condition must first be diagnosed, which can be achieved using EEG.

Like TUAB, TUEP [9] is also an annotated subset of the TUEG corpus. It contains recordings from 100 patients with epilepsy and 100 patients without epilepsy. Despite the equal number of subjects, the number and total length of recordings from epileptic subjects exceeds that of non-epileptic subjects by a factor greater than 4, hence causing a significant class imbalance in the dataset.

A summary of basic statistics for TUEP are given in table 18.

Number of subjects	200
Sessions per subject (average)	10.2
Average session length	18.6 min
EEG channels	19
Sampling rates	$\{250, 256, 400, 512, 1000\}$ Hz
Number of epilepsy recordings	1,651
Number of recordings without epilepsy	390

Table 18: Summary of the TUEP dataset

D.12 TUAR

The TUAR dataset [19] contains annotations of "artifacts" in the EEG data. Artifacts are disturbances in an EEG recording that are not caused by the recorded brain, but rather by external factors. In TUAR, these include movements by the participant, such as chewing, shivers, agitation and eye movements, as well as electrode issues, like interferences or displacements.

As artifacts can be mistaken for other event types, like seizures [19], or can make the detection of an event with which they overlap more difficult, it is often desirable to detect and/or remove artifacts from a given EEG signal before attempting to detect patterns of interest. Hence, the aim of TUAR is aid the development of artifact detection techniques.

TUAR is a subset of the TUEG corpus, containing annotated EEG recordings from 213 patients, taken over 259 sessions. Table 19 presents some basic statistics of the dataset.

D.13 CHB-MIT

CHB-MIT [20] contains annotated seizures from the Children's Hospital Boston. In order to better understand how to counteract their seizures, the 23 subjects had their anti-seizure medication withdrawn for the duration of the study and their brain activity was subsequently recorded. In total, 182 seizure events were annotated. It should be mentioned that the EEG channels are given as a bipolar montage. Hence, none of the foundation models was trained with any of CHB-MIT's channels. In order to still be able to train them on CHB-MIT, we decided to ignore the models' pre-defined channels and use those of CHB-MIT anyway, in an arbitrary (but fixed) order.

A summary of statistics of CHB-MIT can be found in table 20.

Number of subjects	213	
Sessions per subject (average)	1.2	
Average session length	23 min	
EEG channels	19	
Sampling rates	$\{250, 256, 400, 512, 1000\}$ Hz	
Average event length	8 s	
Number of events	Eye Movement	60,577
	Muscle Artifact	81,623
	Electrode Artifact	43,356
	Chewing	7,741
	Shivers	659

Table 19: Summary of the TUAR dataset

Number of subjects	23
Sessions per subject (average)	29.8
Average session length	1.4 h
EEG channels	18
Sampling rate	256 Hz
Average seizure length	62 s
Number of seizures	182

Table 20: Summary of the CHB-MIT dataset

D.14 Sleep-Telemetry

Human sleep can be categorized into different phases, or "stages". Overall, one can distinguish rapid eye movement (REM) phases and non-REM phases, while the non-REM phases can be further differentiated by the "depth" of the sleep (i.e. how hard it is to wake someone up), into N1, N2, N3 and N4 phases (ordered by increasing depth). These phases can be well observed in an EEG recording and hence, this type of measure is used in many areas of the study of sleep.

In this paper, we use the publicly available Sleep-Telemetry [21] dataset, which was part of a study investigating the effects of temazepam medication on sleep behavior. It contains data from 22 subjects aged between 18 and 79 years. For each subject, two nights of nine hours were recorded, one night with temazepam intake, and one night with a placebo. The dataset features three channels: Fpz-Cz, Pz-Oz and horizontal EOG (Electrooculogram). Similar to CHB-MIT, none of the foundation models was trained on any of these channels and hence, we again opt to ignore the model-channels.

An overview of some statistics of the Sleep-Telemetry dataset is given in table 21.

Number of subjects	22	
Sessions per subject	2	
Average session length	8.6 h	
Channels	$2\times$ EEG, $1\times$ EOG	
Sampling rate	100 Hz	
Average event length	3.5 min	
Number of events	Wake Phase	744
	N1 Phase	1311
	N2 Phase	1718
	N3 or N4 Phase	1631
	REM Phase	378

Table 21: Summary of the Sleep-Telemetry dataset

E Dataset Licensing and Usage Compliance

We reviewed and complied with the license or usage terms of all datasets included in EEG-Bench. Below is a dataset-by-dataset breakdown:

- **TUAB**, **TUEP**, **TUAR** – Provided by the Neural Engineering Data Consortium (NEDC) under a public data use agreement.
  We have registered with NEDC and fully comply with their usage terms: (1) the dataset providers are acknowledged as requested, (2) we do not redistribute the data, (3) no attempts are made to re-identify subjects, (4) we use the data solely for research and non-malicious purposes, and (5) we agree to delete the data after use if required.
- **Cavanagh2018a**, **Cavanagh2018b**, **Cavanagh2019**, **Albrecht2019**, **Singh2018**, **Brown2020**, **Gruendler2009**, **Singh2020**, **Singh2021** – Licensed under *Public Domain Dedication and License (PDDL) v1.0*.
  Used in accordance with public domain status for academic research.
- **CHB-MIT**, **Sleep-Telemetry** – Licensed under *Open Data Commons Attribution License v1.0*.
  Used with appropriate attribution for research purposes.

All datasets are used strictly for non-commercial, academic purposes. No data is redistributed or altered in violation of its license, and no attempts at re-identification or deanonymization have been made.