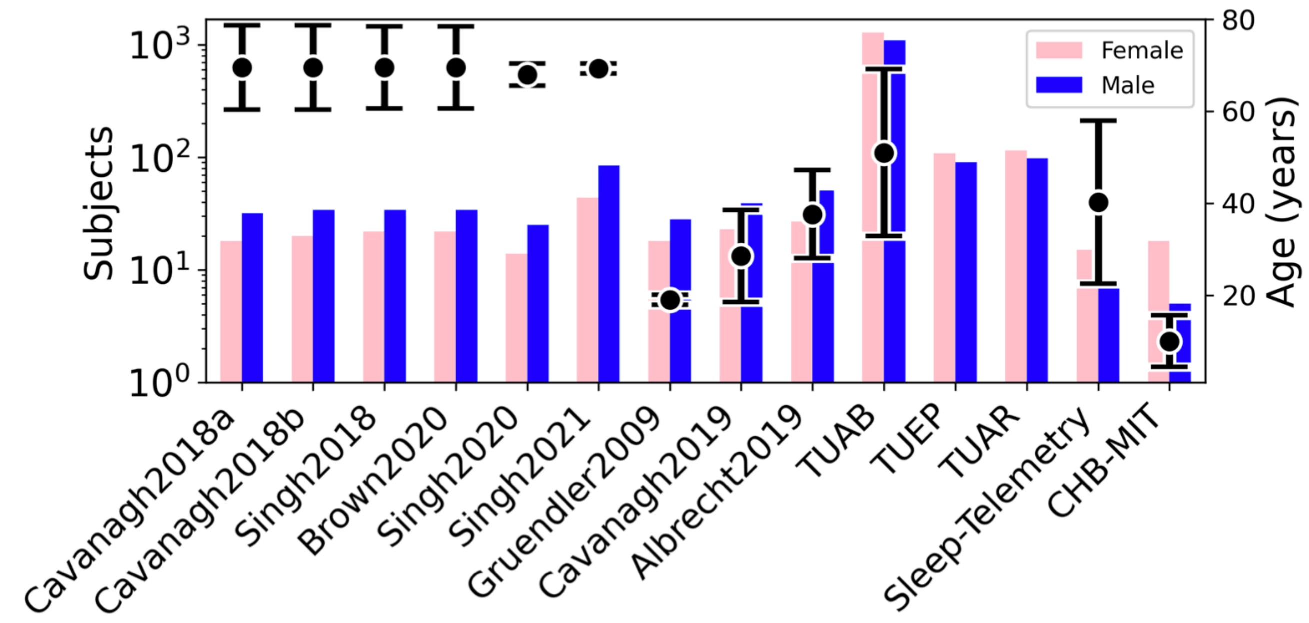
Contact us: akastrati@ethz.ch, jbuerki@ethz.ch, jlauer@ethz.ch, cxuan@ethz.ch, iaquintr@ethz.ch, wattenhofer@ethz.ch

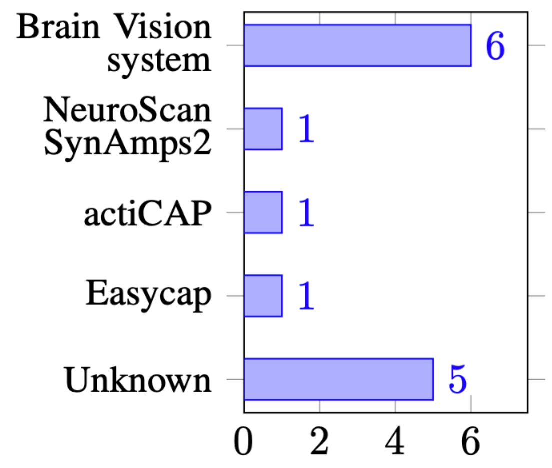
Paper Code Dataset

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

Ard Kastrati, Josua Bürki, Jonas Lauer, Cheng Xuan, Raffaele Iaquinto, Roger Wattenhofer
ETH Zurich, Distributed Computing Group

DATASET





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: x20 stimulus, x3-x5 blocks; 4s	64	31 h	56	84
Gruendler2009 [16]	OCD	Reward learning: 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); 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

Datset Diversity and Experimental Paradimgs: Heterogeneous EEG systems, broad demographics, and varied clinical and cognitive tasks spanning PD, mTBI, OCD, eye-state, sleep, oddball, and motor paradigms.

BENCHMARK

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

Tasks: 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.

MODELS

- Standard ML: SVM, LDA
- Foundation Models: BENDR, Neuro-GPT, LaBraM

RESULTS

- Foundation models excel in some settings, but simpler models often hold up better under real clinical variability.
  - Foundation models excel in some tasks (e.g., LaBraM: abnormal EEG 0.838, OCD 0.740)
  - Classical ML remain competitive in some tasks (e.g., LDA: mTBI 0.813, SVM: schizophrenia 0.679)

EXPERIMENTS

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**

Experiments: 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.

CONCLUSIONS

Our contributions are threefold:

- Standardized & Extensible Framework: Unified parsing; Minimal preprocessing; Classical & Foundation Models support
- Well-defined Clinical Tasks: 11 tasks; Diagnostic & event-based; Neurological & psychiatric conditions
- Comprehensive Performance Evaluation: Baselines vs. foundation models; Cross-subject generalization; Clinical robustness

All data and code are openly available for reproducibility and adoption. 😄