Toward OpenEEG-Bench: A Live Community-Driven Benchmark for EEG Foundation Models

Pierre Guetschel
Donders Institute for Brain, Cognition and Behaviour, Radboud University
Nijmegen, The Netherlands
0000-0002-8933-7640

Bruno Aristimunha
Yneuro and UCSD
Paris, France and San Diego, USA
0000-0001-5258-2995

Dung Truong
University of California San Diego
USA
dutruong@ucsd.edu

Kuntal Kokate
University of California San Diego
USA
kkokate@ucsd.edu

Michael Tangermann*
Donders Institute for Brain, Cognition and Behaviour, Radboud University
Nijmegen, The Netherlands
0000-0001-6729-0290

Arnaud Delorme*
SCCN, INC, SDSC
University of California San Diego, USA
CNRS, France
adelorme@ucsd.edu

***Abstract*—The rapid emergence of foundation models for electroencephalography (EEG) promises to transform brain-computer interfaces and clinical neuroscience. In many cases, however, results reported on foundation models are snapshots in time, which are hard to compare due to heterogeneous evaluation protocols, such as differing pre-processing, datasets, data-splits, finetuning methods etc.**

To mitigate this, a few benchmark papers have recently been proposed that try to standardize the comparison of foundation models. Compared to other domains, however, the EEG field still lacks a continuously updated, open-source benchmark which provides a *live online leaderboard* and allows for a fair, reproducible comparison of current foundation models, and also allows for introspecting which design decisions are impactful.

In this paper, building upon our experience organizing the NeurIPS 2025 EEG Foundation Model Challenge, we describe our four initial design choices for a live EEG foundation model benchmark, that implements a continuously updated leaderboard: (1) It embraces existing open-source tools, including MNE-Python, Braindecode, and HuggingFace for easier adoption. (2) We recommend using openly accessible datasets and will include new datasets in the future. (3) We standardize the finetuning and pre-processing procedures for comparable results. (4) We implement a community-driven governance to ensure long-term sustainability.

To stimulate feedback and discussion about our current design choices, we also present the initial results for five recent foundation models evaluated across ten commonly used datasets. Before we freeze design decisions, we invite the community to share suggestions at https://huggingface.co/spaces/braindecode/EEG-finetune-arena/discussions. The live leaderboard will be accessible at the same location. The code is open-source and can be found at https://github.com/braindecode/OpenEEGBench.

***Index Terms*—EEG, BCI, SSL, FAIR, Electrophysiology, Electroencephalography, Foundation Models, Benchmarking, Leaderboard, Brain-Computer Interfaces, Deep Learning, Self-Supervised Learning, Reproducibility, Open Science**

* Shared last authorship

I. Introduction

Foundation models (FMs) are large-scale neural networks pre-trained on massive amounts of data, typically in a self-supervised manner, that can be adapted to a wide range of downstream tasks with minimal finetuning. This paradigm has revolutionized machine learning across multiple domains. In natural language processing, models such as GPT-3 [1] have demonstrated remarkable capabilities in text understanding and generation. In computer vision, models such as CLIP [2], SAM [3], and V-JEPA [4] have achieved state-of-the-art performance on diverse visual tasks. Similarly, in the audio domain, models like Whisper [5] have pushed the boundaries of speech recognition and audio understanding. The success of foundation models in these domains has enabled previously unattainable capabilities: zero-shot classification and generation without task-specific training [1], [2], cross-modal retrieval between text and images [2], universal segmentation from natural language prompts [3], and multilingual speech recognition from a single unified model [5].

Inspired by these advances, the electroencephalography (EEG) community has witnessed a rapid development of foundation models in recent years [6]–[10]. This progress has been enabled by two main factors: (1) the growing availability of large-scale EEG data sources, including the Temple University Hospital EEG Corpus [11], OpenNeuro [12], NEMAR [13], the Healthy Brain Network dataset [14], and MOABB [15]; and (2) the successful adaptation of self-supervised learning techniques from other modalities—such as masked autoencoders (MAE) [16] and joint-embedding predictive architectures (JEPA) [4]—to EEG data [17], [18], allowing to leverage unlabeled recordings for pre-training. EEG foundation models hold significant promise for advancing

brain-computer interfaces and neural signal analysis. They offer the potential for rapid finetuning to novel tasks, reducing the need for large labeled datasets that are costly and time-consuming to acquire. Furthermore, these models may provide channel-set invariance, which enables transfer learning across different electrode configurations, and robustness to missing or corrupted channels, improving their usability in real-world clinical and consumer applications. Notably, the winning team of the recent EEG Foundation Challenge [19] used a foundation model approach, with self-supervised learning on large EEG datasets followed by finetuning for the challenge's downstream task. This success provides strong evidence for the effectiveness and promise of such models in advancing the field.

However, rigorous evaluation and standardized comparison of these models is essential for the field to advance. The standard procedure for evaluating a foundation model consists of (1) loading the pre-trained weights of the backbone architecture, (2) appending a newly initialized classification head, and (3) finetuning and testing the resulting model on a downstream dataset. Yet comparing results across publications poses significant challenges, as this evaluation process involves numerous design choices that are rarely consistent between studies. These inconsistencies include heterogeneous test datasets with varying preprocessing pipelines, diverse classification heads appended to the backbone architectures, and differing finetuning procedures. This lack of standardization makes it difficult to assess which models truly perform best, or even to attribute performance differences to the backbone architecture itself. Therefore, researchers introducing new foundation models have to re-implement and re-evaluate prior work, which significantly raises the barrier to entry and limiting the number of baselines that can be feasibly compared.

Recent benchmarking efforts [20]–[22] have attempted to address reproducibility concerns, yet they fall short of providing a sustainable solution. While these works offer valuable comparisons, the associated code is not always publicly released, and the benchmarks are typically designed as one-time evaluations accompanying a publication rather than as living resources for ongoing community use. Crucially, no existing project has the mandate or legitimacy to govern evaluation practices across the fragmented EEG software ecosystem. Here, standards and tools have recently emerged that received broad support from the community. Examples are the BIDS (Brain Imaging Data Structure) standard which defines data organization and metadata conventions. Tools like MNE-Python [23] and EEGPrep [24] provide implementations for preprocessing steps. Braindecode [25] provides standardized implementations of deep learning models and foundation models, and the necessary tools to train them. Additionally, the field is converging on a common set of datasets for benchmarking, including TUAB [11], PhysioNet-MI [26], and BCIC-IV-2a [27]. However, there is currently no overarching infrastructure connecting the libraries together and consuming the datasets to build a benchmark. The necessary tools exist; what is missing is community coordination.

In this paper, we describe our ongoing work toward addressing these limitations. Rather than building from scratch, we assemble existing mature open-source tools—MNE-Python for preprocessing, Braindecode for model training, and HuggingFace for hosting—into a unified evaluation framework. A key difference from previous efforts is our **live online leaderboard**, continuously updated as new models are submitted, which transforms the benchmark from a static publication into a living community resource. We envision community-driven governance to ensure long-term sustainability and broad adoption.

Building on our experience organizing the EEG Foundation Challenge [19]—which attracted over 1,000 participants—we present our initial design choices and preliminary results on 10 commonly used datasets. This framework is work-in-progress: we share our current approach to gather early community feedback before finalizing the benchmark.

II. Proposed Benchmark Framework

We propose a benchmark framework consisting of three core components aligned with the requirements outlined above.

A. Benchmark Specification

The benchmark specification defines the community-owned standard for evaluation. It comprises: (i) the set of approved datasets with their preprocessing requirements, (ii) task definitions and evaluation metrics, (iii) cross-validation protocols, and (iv) reporting requirements. This specification is versioned and maintained through a public GitHub repository with a formal change proposal process.

B. Governance and Dataset Selection

We envision community-driven governance for the benchmark, drawing inspiration from successful models used by other scientific software communities such as BIDS, including a steering committee for making decisions. A detailed description of the foreseen governance model is provided in the Supplementary Materials online, though the exact structure remains open for discussion and community input.

For this initial release, we selected 10 datasets which are listed in Table I. These correspond to the most commonly used for benchmarking in recent EEG foundation model publications, as shown in the *Used by* column. This pragmatic approach ensures compatibility with existing literature.

Notably, several of these datasets are not openly accessible (i.e., they require institutional agreements or application-based access), as visible in the *Open* column. We argue that **non-open datasets should be deprecated** from future benchmark versions. While these datasets have historical significance and appear frequently in the literature, their access restrictions fundamentally conflict with reproducibility goals. A benchmark built on restricted data cannot be independently verified, and researchers without institutional access are excluded from participation. As the community governance structure matures, we strongly recommend only adding truly open datasets with permissive licenses that allow redistribution and derivative works in the future and deprecating the use of non-open ones.

TABLE I
INITIAL BENCHMARK DATASETS

Dataset	Paradigm	Subjects	Classes	Used by	Open
PhysioNet-MI [26]	Motor imagery	109	4	[9], [10]	✓
BCIC-IV-2a [27]	Motor imagery	9	4	[8], [10]	✓
BCIC-2020-3 [28]	Imagined speech	25	11	[9], [10]	✓
TUEV [11]	Event classif.	300+	6	[6]–[10]	×
TUAB [11]	Abnormal detect.	2300+	2	[6]–[10]	×
ISRUC [29]	Sleep staging	100	5	[9], [10]	✓
Mumtaz [30]	Mental disorder	64	2	[9], [10]	✓
MAT [31]	Mental stress	36	2	[8], [10]	✓
SEED-V [32]	Emotion recog.	16	5	[8], [9]	×
FACED [33]	Emotion recog.	123	9	[8]–[10]	×

C. Implementation

a) Backend: The **evaluation harness** is the reference implementation that validates model submissions. Built on Braindecode, it provides:

- Automated data loading with standardized preprocessing via MNE-Python/EEGPrep
- Standardized finetuning protocols with fixed classification heads

This evaluation harness is version-fixed and provides an evaluation environment that ensures numerical reproducibility across platforms. It is publicly available and can be used by researchers to evaluate their models locally before submission, ensuring that results are consistent with the online leaderboard.

b) Frontend: A central feature of our framework is the **live online leaderboard**, which distinguishes it from previous one-time benchmark publications. Researchers can submit their models on HuggingFace and receive standardized evaluation results. Under the hood, HuggingFace will automatically run the evaluation harness as new models are submitted to continuously update the leaderboard, always reflecting the current state-of-the-art. The HuggingFace infrastructure additionally provides:

- Dataset hosting with standardized data loaders
- Model hub integration for pre-trained foundation model weights
- Automated evaluation pipelines triggered on model submission
- Off-the-shelf finetuning methods (see subsection III-C)

This transforms the benchmark from a static publication into a living resource that evolves with the field, and eliminates the need for researchers to re-implement prior baselines.

III. PRELIMINARY EXPERIMENTS

The following experiments validate the feasibility of our framework and inform ongoing design decisions. These are not final benchmark scores but initial results to identify open questions and gather feedback.

Importantly, **our goal is not to exactly reproduce results reported in original publications**. Rather, we aim to provide a framework that enables fair comparison of pretrained backbones by decoupling the many moving parts that are currently entangled across studies: preprocessing pipelines, classification heads, finetuning procedures, dataset splits, and evaluation metrics. When all these aspects vary simultaneously, it becomes impossible to attribute performance differences to the backbone architecture itself.

A. Preprocessing

We apply minimal preprocessing to preserve signal characteristics while ensuring compatibility across datasets. All recordings are high-pass filtered at 0.1 Hz, except for tasks with short trial windows (2 s or less), where a 0.5 Hz cutoff is used to avoid excessive filter transients.

Subsequently, we apply normalization to the preprocessed signals. The specific normalization method (z-score standardization, etc.) varies depending on the requirements of each foundation model, as detailed in the *Normalisation* column of Table III.

B. Data split

We employ predefined train/validation/test splits to assess generalization to unseen subjects, following practices established in recent foundation model papers [7], [9], [10]. Subjects are stratified across splits to ensure balanced representation of data characteristics, except for datasets SEED-V and BCIC-2020-3 where task difficulty would make strict cross-subject transfer impossible; instead, we used the within-session split provided in the original publications. The exact split details are provided in the Code and Supplementary Materials Online.

C. Training

a) Finetuning methods and classification heads: A core challenge in comparing foundation models is that each publication uses different classification heads and finetuning procedures, making it difficult to isolate the backbone's contribution. To decouple model evaluation from classification head design, we use a single linear layer that receives flattened features from the backbone. We evaluate two finetuning approaches: (1) **Probe**—freezing the backbone and training only the classification head, and (2) **Full finetuning**—training both backbone and head end-to-end. While simpler than methods like those in REVE [10], these approaches enable fair comparisons across models.

b) Training hyperparameters: Models are finetuned with the AdamW optimizer (learning rate 5×10^{-4}, weight decay 0.01, $\beta = (0.9, 0.999)$), cosine annealing schedule with linear warmup over 5 epochs, and gradient clipping (max norm 1.0). Early stopping monitors the validation loss with patience of 10 epochs (minimum improvement threshold 0.001). The best checkpoint is selected based on validation accuracy.

D. Results

For this initial evaluation, we benchmark five recent EEG foundation models, listed in Table III. The results are presented in Table II, where the balanced accuracy score was linearly normalized to [0,100], mapping chance level to 0 % and perfect performance to 100 %. These initial results should not be

TABLE II
BENCHMARK RESULTS AND RESULTS REPORTED BY THE AUTHORS

Model	Dataset / Fine-tuning Method	PhysioNet-MI	BCIC-IV-2a	BCIC-2020-3	TUEV	TUAB	ISRUC	FACED	Mumtaz	MAT	SEED-V	Average
BIOT	Linear Probing	1.71	1.83	2.03	22.78	14.53	62.74	4.95	64.97	10.52	4.82	19.09
	Full Finetuning	3.97	5.75	2.26	31.14	-1.01	69.63	6.90	63.31	**34.77**	12.73	22.94
LaBraM	Linear Probing	5.85	1.03	0.05	6.52	20.41	33.21	1.57	56.52	-0.25	3.55	12.85
	Full Finetuning	42.38	14.80	1.53	25.98	54.18	69.45	5.85	71.56	11.11	**15.54**	31.24
	Reported by authors	-	-	-	*59.39*	*65.16*	-	-	-	-	-	*62.28*
EEGPT	Linear Probing	8.22	6.20	0.82	20.57	42.98	35.92	7.09	55.50	10.14	1.53	18.90
	Full Finetuning	39.04	13.89	0.60	6.71	40.67	62.11	14.49	66.93	12.50	1.32	25.83
	Reported by authors	-	-	-	-	-	-	*9.89*	-	-	*17.38*	*13.63*
CBraMod	Linear Probing	27.13	4.70	2.28	13.05	10.05	24.90	1.21	55.80	-0.25	1.31	14.02
	Full Finetuning	43.52	31.06	5.41	35.69	58.37	**72.15**	5.66	**74.09**	26.44	8.71	36.11
	Reported by authors	*52.23*	-	-	*60.05*	*63.60*	-	*49.48*	-	-	*26.14*	*50.30*
REVE	Linear Probing	37.67	19.11	**8.84**	37.91	-	64.64	16.72	68.11	29.54	7.63	32.24
	Full Finetuning	**43.60**	**35.19**	4.06	**38.45**	**61.12**	61.96	**18.28**	67.51	32.27	6.62	**36.91**
	Reported by authors	*53.07*	*51.95*	*45.44*	*61.11*	*66.30*	*72.74*	*51.02*	*92.88*	*53.20*	-	*60.85*

TABLE III
FOUNDATION MODELS EVALUATED

Model	**Publication year**	**Normalization**
BIOT [6]	2023	Per-channel, per-window scaling to 95$^{\text{th}}$ percentile.
LaBraM [7]	2024	Divide signals (µV) by 100.
EEGPT [8]	2024	Global average reference and set unit to 1 mV.
CBraMod [9]	2025	Divide signals (µV) by 100.
REVE [10]	2025	Session-wise scaling and clipping over 15 std.

interpreted as a definitive ranking of these models as the scores may evolve as we integrate the feedback we will gather from the community to the evaluation harness. Nevertheless, we can already make the following observations.

a) Most architectures require full finetuning: When using the linear probing method, all models except REVE stay around chance-level performance (0.0) on at least one dataset. This indicates that their learned representations are not linearly separable for those tasks. This finding suggests that current foundation models do not yet fulfill a key promise of the foundation model paradigm: the ability to easily adapt to new tasks through lightweight finetuning of only a linear classifier.

b) Large gap with reported results, as expected: We observe substantial differences between our benchmark scores and those reported in the original publications. This gap was expected, as the original papers employ heterogeneous preprocessing pipelines, classification heads, and finetuning procedures. These discrepancies underscore the critical need for decoupling the evaluation of backbone architectures from other design choices, which was a primary objective of our benchmark framework.

IV. CONCLUSION AND CALL FOR FEEDBACK

We have outlined our vision and initial progress toward a community-driven benchmark for EEG foundation models. By assembling existing open-source tools into a unified framework with a live online leaderboard, we aim to transform evaluation from one-time publication artifacts into a continuously updated resource.

This work is in progress. We share our current design choices and preliminary results to gather early community feedback before the code release and online launch on:

- Open-source software infrastructure and compatibility with existing tools
- Dataset selection and coverage
- Preprocessing and evaluation protocols
- Governance structure and contribution processes
- Leaderboard features and submission workflows

We invite the community to contribute feedback and suggestions via the *Community* tab on our HuggingFace space at https://huggingface.co/spaces/braindecode/EEG-finetune-arena. The online leaderboard will be made available at the same location. The code is open-source and can be found at https://github.com/braindecode/OpenEEGBench.

ACKNOWLEDGMENTS

We acknowledge Amitava Majumdar for his leadership of the Neuroscience Gateway (NSG) project at the San Diego Supercomputer Center, which provides community access to high performance computing resources for large scale neuroscience research. Funding source for GPU allocation: NAIRR250045.

REFERENCES

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 1877–1901. [Online]. Available: https://papers.nips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html

[4] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," *arXiv preprint arXiv:2404.08471*, 2024. [Online]. Available: https://arxiv.org/abs/2404.08471

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[6] C. Yang, M. B. Westover, and J. Sun, "BIOT: Biosignal transformer for cross-data learning in the wild," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference.html

[7] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu, "Large brain model for learning generic representations with tremendous EEG data in BCI," in *International Conference on Learning Representations (ICLR)*, 2024. [Online]. Available: https://arxiv.org/abs/2405.18765

[8] G. Wang, W. Liu, Y. He, C. Xu, L. Ma, and H. Li, "EEGPT: Pretrained transformer for universal and reliable representation of EEG signals," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: https://openreview.net/forum?id=lvS2b8CjG5

[9] J. Wang, S. Zhao, Z. Luo, Y. Zhou, H. Jiang, S. Li, T. Li, and G. Pan, "CBraMod: A criss-cross brain foundation model for EEG decoding," in *International Conference on Learning Representations (ICLR)*, 2025. [Online]. Available: https://arxiv.org/abs/2412.07236

[10] Y. El Ouahidi, J. Lys, P. Thölke, N. Farrugia, B. Pasdeloup, V. Gripon, K. Jerbi, and G. Lioi, "REVE: A foundation model for EEG: Adapting to any setup with large-scale pretraining on 25,000 subjects," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. [Online]. Available: https://brain-bzh.github.io/reve/

[11] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers in Neuroscience*, vol. 10, p. 196, 2016. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2016.00196/full

[12] C. J. Markiewicz, K. J. Gorgolewski, F. Feingold, R. Blair, Y. O. Halchenko, E. Miller, N. Hardcastle, J. Wexler, O. Esteban, M. Goncavles, A. Jwa, and R. Poldrack, "The OpenNeuro resource for sharing of neuroscience data," *eLife*, vol. 10, p. e71774, 2021. [Online]. Available: https://elifesciences.org/articles/71774

[13] A. Delorme and S. Makeig, "NEMAR: NeuroElectroMagnetic data archive," https://nemar.org/, 2024, accessed: 2025.

[14] S. Y. Shirazi, A. Franco, M. Scopel Hoffmann, N. Esper, D. Truong, A. Delorme, M. Milham, and S. Makeig, "HBN-EEG: The FAIR implementation of the healthy brain network (HBN) electroencephalography dataset," *bioRxiv*, 2024. [Online]. Available: https://www.biorxiv.org/content/10.1101/2024.10.03.615261

[15] B. Aristimunha, I. Carrara, P. Guetschel, S. Sedlar, P. Rodrigues, J. Sosulski, D. Narayanan, E. Bjareholt, Q. Barthelemy, R. T. Schirrmeister, R. Kobler, E. Kalunga, L. Darmet, C. Gregoire, A. Abdul Hussain, R. Gatti, V. Goncharenko, J. Thielen, T. Moreau, Y. Roy, V. Jayaram, A. Barachant, and S. Chevallier, "Mother of all bci benchmarks," Nov. 2025. [Online]. Available: https://doi.org/10.5281/zenodo.10034223

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 000–16 009. [Online]. Available: https://arxiv.org/abs/2111.06377

[17] H.-Y. S. Chien, H. Goh, C. M. Sandino, and J. Y. Cheng, "MAEEG: Masked auto-encoder for EEG representation learning," in *NeurIPS Workshop on Learning from Time Series for Health*, 2022. [Online]. Available: https://arxiv.org/abs/2211.02625

[18] P. Guetschel, T. Moreau, and M. Tangermann, "S-jepa: towards seamless cross-dataset transfer through dynamic spatial attention," in *9th Graz Brain-Computer Interface Conference*, Graz, Austria, September 2024. [Online]. Available: https://arxiv.org/abs/2403.11772

[19] B. Aristimunha, D. Truong, P. Guetschel, S. Y. Shirazi, I. Guyon, A. R. Franco, M. P. Milham, A. Dotan, S. Makeig, and A. Delorme, "EEG foundation challenge: From cross-task to cross-subject EEG decoding," in *The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025. [Online]. Available: https://arxiv.org/abs/2506.19141

[20] Anonymous, "Are EEG foundation models worth it? comparative evaluation with traditional decoders in diverse BCI tasks," in *Submitted to International Conference on Learning Representations*, 2026, under review. [Online]. Available: https://openreview.net/forum?id=5Xwm8e6vbh

[21] G. Kuruppu, N. Wagh, V. Kremen, S. Pati, G. Worrell, and Y. Varatharajah, "EEG foundation models: A critical review of current progress and future directions," *arXiv preprint arXiv:2507.11783*, 2025. [Online]. Available: https://arxiv.org/abs/2507.11783

[22] J. Lai, J. Wei, L. Yao, and Y. Wang, "A simple review of EEG foundation models: Datasets, advancements and future perspectives," *arXiv preprint arXiv:2504.20069*, 2025. [Online]. Available: https://arxiv.org/abs/2504.20069

[23] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hämäläinen, "MEG and EEG data analysis with MNE-Python," *Frontiers in Neuroscience*, vol. 7, p. 267, 2013. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2013.00267/full

[24] EEGPrep Contributors, "EEGPrep: Python eeg preprocessing pipeline reproducing the eeglab default preprocessing workflow," 2025, accessed: 2026-02-09. [Online]. Available: https://github.com/sccn/eegprep

[25] B. Aristimunha, P. Guetschel, M. Wimpff, L. Gemein, C. Rommel, H. Banville, M. Sliwowski, D. Wilson, S. Brandt, T. Gnassounou, J. Paillard, B. Junqueira Lopes, S. Sedlar, T. Moreau, S. Chevallier, A. Gramfort, and R. T. Schirrmeister, "Braindecode: toolbox for decoding raw electrophysiological brain data with deep learning models," 2025. [Online]. Available: https://doi.org/10.5281/zenodo.8214376

[26] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000. [Online]. Available: https://www.physionet.org/

[27] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI competition IV," *Frontiers in Neuroscience*, vol. 6, p. 55, 2012. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2012.00055/full

[28] J.-H. Jeong, J.-H. Cho, K.-H. Shim, B.-H. Kwon, B.-H. Lee, D.-Y. Lee, D.-H. Lee, and S.-W. Lee, "Multimodal signal dataset for 11 intuitive movement tasks from single upper extremity during multiple recording sessions," *GigaScience*, vol. 9, no. 10, p. giaa098, 2020.

[29] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-Sleep: A comprehensive public dataset for sleep researchers," *Computer Methods and Programs in Biomedicine*, vol. 124, pp. 180–192, 2016.

[30] W. Mumtaz, "MDD Patients and Healthy Controls EEG Data (New)," 11 2016. [Online]. Available: https://figshare.com/articles/dataset/EEG_Data_New/4244171

[31] I. Zyma, S. Tukaev, I. Seleznov, K. Kiyono, A. Popov, M. Chernykh, and O. Shpenkov, "EEG during mental arithmetic tasks," *Data*, vol. 4, no. 1, p. 14, 2019. [Online]. Available: https://www.mdpi.com/2306-5729/4/1/14

[32] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715–729, 2022.

[33] C. Chen, Z. Li, K. Yu, H. Zhu, B.-L. Wang, and W. Chen, "FACED: A finer-grained affective computing EEG dataset for emotion recognition," *Scientific Data*, vol. 10, p. 737, 2023.