ASSESSING THE CAPABILITIES OF LARGE BRAINWAVE FOUNDATION MODELS

Na Lee*
Imperial College London & Cogitat
na.lee12@imperial.ac.uk

Stylianos Bakas*
Aristotle University of Thessaloniki &
Imperial College London & Cogitat

Konstantinos Barmpas
Imperial College London &
Archimedes / Athena RU & Cogitat

Yannis Panagakis
National and Kapodistrian University of Athens &
Archimedes / Athena RU & Cogitat

Dimitrios A. Adamos
Imperial College London & Cogitat

Nikolaos Laskaris
Aristotle University of Thessaloniki & Cogitat

Stefanos Zafeiriou
Imperial College London & Cogitat

ABSTRACT

Over the last decade, deep learning models have been widely used for automatic feature extraction and classification in various Brain-Computer Interface (BCI) tasks. However, their performance and generalization capabilities are often not adequately assessed, as these models are frequently trained and tested under flawed setups and / or influenced by spurious correlations. Recently, these limitations have also been observed in the training and evaluation of Large Brainwave Foundation Models (LBMs). In this work, we employ causal reasoning and careful consideration for task-discriminative artifacts in various EEG datasets covering diverse BCI paradigms and propose a benchmarking protocol to properly evaluate the decoding performance and generalization capabilities of LBMs. Utilising a subject-independent cross-validation approach for each curated benchmark dataset, we showcase that LBMs achieve marginal performance gains over conventional deep learning baselines.

1 INTRODUCTION

Brain-Computer Interfaces (BCIs) promise a new way to interact with machines by establishing direct communication between the human brain and computers. This technology is based on the analysis of brainwaves from electroencephalogram (EEG) recordings and finds application in various areas like emotion recognition Torres et al. (2020); Xu et al. (2018), epileptic seizure detection Alkawadri (2019) and motor-imagery Rezaeitabar & Halici (2017). BCIs can also augment human abilities and have the potential to transform how we interact with our environment and each other, offering hope to people with disabilities to regain lost functions Chaudhary et al. (2016); Luu et al. (2017); Biasiucci et al. (2018).

Human experts and manual feature extraction were at the center of of brainwave analysis for many years Bashashati et al. (2007); Handy (2009); McFarland et al. (2006). Recently however, extensive research has been conducted to replace them with data-driven deep learning models Lawhern et al. (2018); Santamaría-Vázquez et al. (2020); Song et al. (2023); Barmpas et al. (2023a); Bakas et al. (2022). Although deep learning models have demonstrated impressive results, they generally require substantial supervision and task-specific data collection, making the process both time-consuming and resource-intensive. Inspired by the tremendous progress of Foundation Models in various fields of Computer Science, Large Brainwave Foundation Models (LBMs) in the domain of BCIs have

*These authors contributed equally to this work.

been recently introduced Jiang et al. (2024); Cui et al. (2024). These large models leverage extensive self-supervised pre-training on diverse unlabeled EEG datasets in order to identify complex patterns. This enables them to generalize effectively and reduce the need for task-specific data collection. The increased generalisation potential fuels the expectation that LBMs will be more robust and adaptable to different users, tasks and environments Barmpas et al. (2024a).

Despite the tremendous progress in the development of brainwave decoding architectures, there is a lack of understanding how LBMs could be effectively used in brainwave decoding and how properly compare these models. Faulty evaluation setups could potentially misrepresent the model performance while their reliance on task-discriminative artifacts (e.g. eye activity) hinders their generalization capabilities. In this work:

1. Building on previous works Barmpas et al. (2024b) and Barmpas et al. (2024a) in causal reasoning within the domain of BCIs and incorporating insights from Bakas et al. (2025) regarding task-discriminative artifacts, we propose a benchmarking protocol designed to evaluate both the performance and generalization capabilities of LBMs as well as other traditional deep learning architectures for brainwave decoding.
2. Using this benchmarking protocol, we evaluate the performance of state-of-the-art LBMs against standard deep learning models for brainwave decoding, providing a comprehensive analysis of their strengths and limitations.

2 PREMILINARIES

2.1 CAUSAL REASONING IN BCIS

Causal reasoning is the analysis of a task/problem in terms of cause-effect relationships between the different variables of interest Castro et al. (2020): if a variable A is a direct cause of variable B, we express it as $A \rightarrow B$. The work in Barmpas et al. (2024b) introduces a causal framework to analyze BCI paradigms by decomposing them into independent variables, following the Principle of Independent Causal Mechanism Reichenbach (1956). The framework relies on two core properties:

- **The presence of task stimulus**: if the brain activity is modulated by the presence of external stimuli (Exogenous) or without any external stimulus (Endogenous)
- **The voluntary engagement of the subject**: when the subject willingly generates a particular brain activation or coactivation pattern (Voluntarily Engaged) or when the subject has no control of the generated brain activation pattern (Involuntarily Engaged)

According to Barmpas et al. (2024b), this categorization allows all BCI paradigms to be classified based on these factors. The identified variables of interests in a BCI task are: the observed EEG brainwave signal X, the labelled task Y, the true underlying unobserved BCI-relevant brain activity Z, the stimulus S, the intention I of the subject and the environment E. For each category, there is a specific directed acyclic graph (DAG) that represents the causal relations between the identified variables of interests:

- Voluntarily Engaged - Exogenous: $\mathsf{Y} \rightarrow \mathsf{S} \rightarrow \mathsf{Z} \rightarrow \mathsf{X}$
- Involuntarily Engaged - Exogenous: $(\mathsf{Y}, \mathsf{S}) \rightarrow \mathsf{Z} \rightarrow \mathsf{X}$
- Involuntarily Engaged - Endogenous: $\mathsf{E} \rightarrow \mathsf{Y} \rightarrow \mathsf{Z} \rightarrow \mathsf{X}$
- Voluntarily Engaged - Endogenous: $\mathsf{I} \rightarrow \mathsf{Y} \rightarrow \mathsf{Z} \rightarrow \mathsf{X}$

Distribution shifts affecting brainwave decoding systems can arise from (see Appendix A):

- **Experimental settings** that leads to shifts in stimuli ($P(\mathsf{S})$), intentions ($P(\mathsf{I})$) or environmental conditions ($P(\mathsf{E})$)
- **EEG data quality and acquisition** that affects the relationship between brain activity and EEG signals ($P(\mathsf{X}|\mathsf{Z})$)
- **Subject variability** that impacts brain activity distributions ($P(\mathsf{Z}|\cdot)$)
- **Class imbalances** in task labels ($P(\mathsf{Y})$)

Each type of shift influences specific causal sub-modules and can negatively impact model performance. Understanding these shifts is essential for designing robust models for brainwave analysis.

2.2 SIGNAL ARTIFACTS

Signal artifacts, such as ocular activity, pose a significant challenge in decoding brainwaves, as they can substantially affect model performance. These artifacts not only reduce the signal-to-noise ratio but can also become class-informative, causing models to overfit to spurious patterns in the data rather than the true neural signals Frølich et al. (2015). This issue is particularly problematic in EEG-based brainwave decoding, where noise from eye movements, blinks and other physiological sources is extensive.

Artifact removal techniques, such as Independent Component Analysis (ICA), have been shown to mitigate the impact of these unwanted signals Urigüen & Garcia-Zapirain (2015); Jiang et al. (2019). However, a large portion of the deep learning brainwave models fails to incorporate these preprocessing steps into their architectural designs Craik et al. (2019). This omission can lead to artificially inflated model performance, as demonstrated in Bakas et al. (2025), where the choice of input window was shown to significantly influence results by capturing class-informative artifacts. Without careful consideration, both deep learning and large models risk overfitting to artifacts or unintended brain signals, compromising their reliability and generalizability in practical applications.

3 PROPOSED PROTOCOL

3.1 CAUSAL CONSIDERATIONS

In this section, taking into consideration the insights from the causal framework as well as artifact concerns described in section 2, we propose a benchmarking protocol for both deep learning models and LBMs for brainwave decoding.

Assessing the generalization capabilities of brainwave models coincides with the subject activity shift ($P(\mathsf{Z}|\cdot)$). As it is highlighted in previous works Barmpas et al. (2023b; 2022), a proper evaluation setup keeps all the identified distribution shifts intact but inter-subject variability, unlike other works which claim improved generalized performance and often utilize a mixture of techniques like data augmentation (which can affect also other causal variables of interest). Therefore, in our benchmarking protocol:

1. Subject-independent cross-validation is utilized for each benchmark dataset
2. Each benchmark dataset contains trials recorded from a single EEG device, preserving $P(\mathsf{X}|\mathsf{Z})$. Additionally, whenever task-feasible, we selected class-balanced datasets, maintaining $P(\mathsf{Y})$

3.2 ARTIFACT CONSIDERATIONS

Choosing the right benchmark datasets is also crucial. The rationale between the selection of the benchmark datasets is twofold:

1. Select downstream tasks that capture a diverse range of BCI paradigms
2. Select datasets with control trials and minimal task-discriminative artifacts

While the first objective is led primarily by the trends in BCI applications (Table 2 offers a wide range of popular downstream BCI tasks), the second objective demands deeper consideration. Inspired by Bakas et al. (2025) our goal is to include datasets with minimal class-discriminative artifacts. In this way, we can ensure that the model is classifying meaningful task-related brain activity and not overfitting on background dynamics or artifacts due to the experimental design. Control trials should also be preferred over inter-trial resting states to minimise ocular artifacts and evoked responses caused by the presented stimulus. In this subsection, we will elaborate on the selection of a dataset that captures a subject's movement using EEG signals. However, similar considerations have been applied to the other benchmark datasets as well.

Capturing a subject's movement using EEG data is one of the most widely used paradigms in Brain-Computer Interface (BCI) research. To select an appropriate benchmark dataset, we first analyzed potential candidates, namely High Gamma dataset Schirrmeister et al. (2017) and PhysioNet dataset Schalk et al. (2004), for the presence of discriminative artifacts, using the methodology of averaged EEG signals and the spatial filters of trained models described in Bakas et al. (2025).

To this end, we first extracted averaged EEG signals on selected electrodes potentially capturing occular artifacts. The brainwave signals were filtered using a bandpass filter in the 1-8Hz frequency range. As it is shown in Figure 1, there is a clear artifact difference between the task-classes during the first second of the trial after the cue appears on the screen on the PhysioNet dataset, with a distinct reverse polarity of the eye movement artifact in F7 and F8 between the left and right hand trials. These clear differences hinder the model's training in the first second, increasing the risk of overfitting and causing it to miss crucial motor-related dynamics at the onset of movement. In contrast, no major time-locked differences between classes can be observed at any point during the trials in the selected electrodes on the High Gamma dataset.

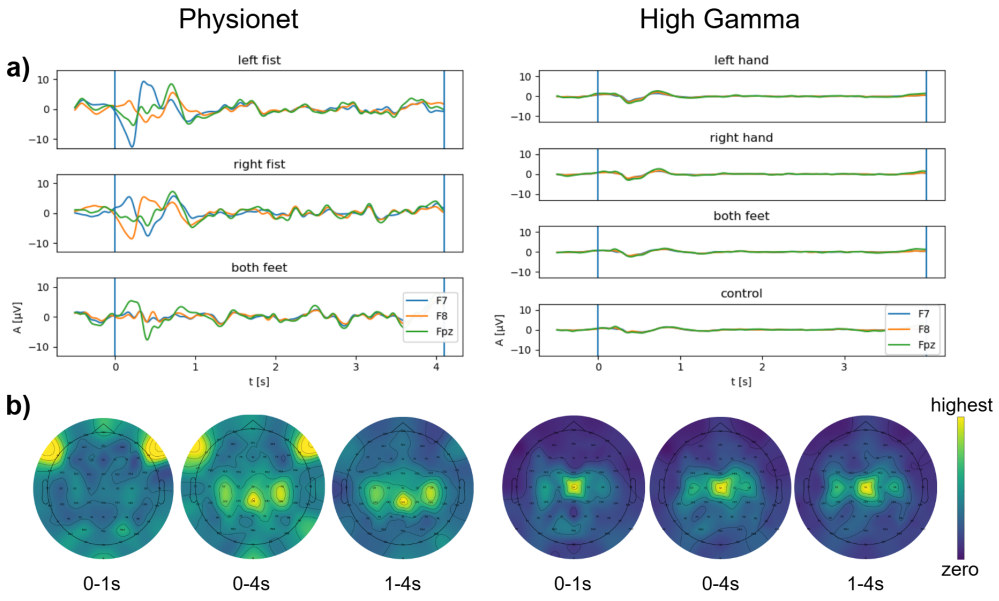

Figure 1: Artifact investigation in motor execution trials in the PhysioNet and High Gamma datasets. a) Grand average EEG amplitude across subjects and trials for frontal electrodes on PhysioNet (left column) and High Gamma (right column). For the PhysioNet dataset, distinct differences in eye activity between the classes are clearly observable in the first second of the trial following the appearance of the cue on the screen. For the High Gamma dataset, no major differences between the classes due to eye activity can be observed at any point during the trial. b) Topographic representation of the respective trained spatial filter layer for classifying motor execution in PhysioNet(left column) and High Gamma (right column), with models trained on different parts of the trial. For PhysioNet, models trained on the first second focus on the F7, F8 and Fpz electrodes, indicating overfitting on occular artifacts caused by the cue presentation. Models trained on the rest of the trial focus on the expected motor-related central electrodes. For High Gamma, models focus on motor-related central electrodes regardless of the selected part of the trial. Spatial filter representations show mean absolute values of the filter weights averaged across filter channels and cross-validation folds to display the electrode relevance.

To further ensure the lack of exploitable artifacts in the selected dataset, we employed the well-established deep learning model EEGNet Lawhern et al. (2018) to classify motor execution based on different parts of the trials and examined the resulting trained spatial filters. The model was trained for 20 epochs while the input signals were filtered using a bandpass filter in the 4-40 Hz frequency range (hyper-parameters were selected based on the original implementation). Figure 1 shows the averaged across folds and filter channels learned spatial filters for models trained on

the first second of the trial, the whole trial and excluding the first second for the selected High Gamma (right column) and PhysioNet (left column) datasets. Models trained on the first second in the PhysioNet dataset place a focus on F7, F8 and Fpz electrodes confirming the class relevant information of the eye activity artifacts. Models trained on the remaining uncontaminated part of the trial focus on the expected central electrodes over the motor cortex. In direct contrast, models trained on the High Gamma dataset focus primarily on central electrodes regardless of the selected input time window affirming the lack of eye-related class-discriminative artifacts in the dataset.

In summary, the design principles of our proposed benchmark protocol, which take into account causal reasoning and artifact considerations, can be summarized in the following step-by-step guidelines.

Table 1: Step-by-step guidelines for designing benchmarks to properly evaluate both the performance and generalization capabilities of brainwave decoders

1. Choose a diverse range of downstream BCI tasks
2. Select a dataset for each downstream BCI task with control trials and assess the existence or absence of artifacts to avoid spurious performance correlations (Artifact Considerations)
3. Accounting for the causal factorization of each task, perform a subject-independent cross-validation for each benchmark task (Causal Reasoning)

In this work, following the above mentioned guidelines, we utilize the following benchmark BCI datasets, shown in Table 2.

Table 2: Selected Benchmark Datasets in our Proposed Protocol

Dataset	Paradigm	Number of Classes
High Gamma Schirrmeister et al. (2017)	Executed Movement	4
OpenBMI-ERP Hong-Kyung et al. (2019)	ERP	2
Pavlov 2022 Pavlov et al. (2022)	Working Memory	2
Sleep-EDF Kemp et al. (2000)	Sleep	6
PhysioNet Schalk et al. (2004)	Eyes Open-Closed	2

4 Experiments

Using the proposed benchmarking protocol, we evaluated the performance of state-of-the-art LBMs, specifically LaBraM Jiang et al. (2024) and NeuroGPT Cui et al. (2024), against standard deep learning models, namely EEGNet Lawhern et al. (2018) and EEGInception Santamaría-Vázquez et al. (2020). This comparison aims to highlight the advantages and limitations of LBMs in comparison to well-established techniques.

For each of the baseline Large Brainwave Models, benchmark data was preprocessed to match the input data structure used during pre-training:

1. For the LaBraM model, a sampling frequency of 200Hz was used, along with a bandpass filter between 0.5-45Hz. Notch filters were applied at 50Hz, 60Hz, and 100Hz to eliminate powerline noise. The trials were divided into 1-second segments, producing 256 patches per sample across all channels. In addition to the EEG trial data, each sample included temporal and spatial embeddings. Temporal embeddings represent the patch's position within the trial's duration, while spatial embeddings indicate the channel's position relative to a global electrode list. Only data from electrodes in the international 10-20 system are used.
2. For the NeuroGPT model, the data were resampled to 250Hz, with a 0.05-100Hz bandpass filter applied. Similar to LaBraM, notch filters at 50Hz, 60Hz, and their harmonics were used. NeuroGPT requires input data to consist of a specific set of channels in a fixed order. Consequently, for each benchmark dataset, only the channels included in the pre-training data were used. If any expected channels were missing, data from the nearest available

electrode (within a few centimeters) was used, and for any missing channels not within range, the data were set to zero.

For each of the baseline LBM, Common Average Re-referencing (CAR) was applied across all channels to reduce noise. In addition, un-trained classification heads were added to the pre-trained LBMs during the finetuning step. The size and structure of the classifier depend on the latent dimension of the model and the number of target classes on the given benchmark. The exact architecture of the classifiers is given in Table 3:

Table 3: Classification heads for each model configuration where n_cls is the number of classes for each benchmark task

MODEL	CLASSIFIER
LABRAM	Linear (200, n_cls), Dropout (0.5)
NEUROGPT FULL MODEL	Linear (1024,256), ELU, Dropout (0.5), Linear (256,32), ELU, Dropout (0.3), Linear (32, n_cls)
NEUROGPT ENCODER	Linear (2160,256), ELU, Dropout (0.5), Linear (256,32), ELU, Linear (32, n_cls)

We finetuned three LBM configurations: the pre-trained LaBraM base model, the pre-trained NeuroGPT model, and the encoder-only module of the pre-trained NeuroGPT model (as suggested by the authors in Cui et al. (2024)). Each configuration was trained for 20 epochs to prevent overfitting and evaluated using 10-fold subject-independent cross-validation as described in Section 3.1. For fair comparison, using the same fine-tuning setup, we performed a similar training (from scratch) process for the EEGNet and EEGInception models to provide a basis for comparison.

Table 4: Classification accuracy of finetuned foundation models and classic architectures. Each trained/finetuned for 20 epochs with 10 fold cross-validation.

MODEL (TRAINABLE PARAMETERS)	MOVEMENT	ERP	MEMORY	SLEEP	EYES	MEAN ACCURACY
EEGNET (2,394)	0.657	**0.912**	<u>0.660</u>	0.624	0.803	0.731
EEGINCEPTION (22,366)	0.590	0.896	**0.669**	<u>0.688</u>	0.823	0.733
LABRAM (5,854,288)	0.614	<u>0.911</u>	0.643	**0.704**	<u>0.840</u>	<u>0.742</u>
NEUROGPT (FULL MODEL) (78,536,146)	<u>0.682</u>	0.904	0.610	0.665	0.821	0.736
NEUROGPT (ENCODER) (717,958)	**0.695**	0.908	0.634	0.647	**0.843**	**0.745**

Table 4 highlights that standard deep learning baselines can match or outperform some large models on specific benchmark tasks. In addition, it is evident that the performance margin over the next-best baseline (which has considerable less trainable parameters) is about 1.0%, indicating that while LBMs show promise, their full potential over standard methods remains unrealized under the proposed benchmarking protocol.

5 CONCLUSION

In this work, we propose a benchmarking protocol based on careful considerations for task-discriminative artifacts and causal reasoning principles. To the best of our knowledge, this is the first work that proposes such a framework that captures a diverse range of BCI paradigms and is set properly to evaluate the performance and generalization capabilities of the recently introduced LBMs. We hope that the proposed benchmarking protocol will be adapted by the research community to assess the capabilities of LBMs. Using this benchmarking protocol, we also evaluated

the performance of state-of-the-art LBMs against standard deep learning models showcasing that although there are marginal performance gains, LBMs have yet to show their promised substantial benefits.

ACKNOWLEDGMENTS

This work was supported by the EPSRC Turing AI Fellowship (Grant Ref: EP/Z534699/1): Generative Machine Learning Models for Data of Arbitrary Underlying Geometry (MAGAL).

REFERENCES

Rafeed Alkawadri. Brain–computer interface (bci) applications in mapping of epileptic brain networks based on intracranial-eeg: An update. *Frontiers in Neuroscience*, 13:191, 2019. ISSN 1662-453X. doi: 10.3389/fnins.2019.00191.

Stylianos Bakas, Siegfried Ludwig, Konstantinos Barmpas, Mehdi Bahri, Yannis Panagakis, Nikolaos Laskaris, Dimitrios A. Adamos, and Stefanos Zafeiriou. Team cogitat at neurips 2021: Benchmarks for eeg transfer learning competition, 2022.

Stylianos Bakas, Siegfried Ludwig, Dimitrios A Adamos, Nikolaos Laskaris, Yannis Panagakis, and Stefanos Zafeiriou. Latent alignment in deep learning models for eeg decoding. *Journal of Neural Engineering*, 2025.

Konstantinos Barmpas, Yannis Panagakis, Dimitrios A. Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A CAUSAL VIEWPOINT ON MOTOR-IMAGERY BRAINWAVE DECODING. *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

Konstantinos Barmpas, Yannis Panagakis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Brainwave-scattering net: a lightweight network for eeg-based motor imagery recognition. *Journal of Neural Engineering*, 20(5):056014, September 2023a. ISSN 1741-2552.

Konstantinos Barmpas, Yannis Panagakis, Stylianos Bakas, Dimitrios A. Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. Improving generalization of cnn-based motor-imagery eeg decoders via dynamic convolutions. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1997–2005, 2023b. doi: 10.1109/TNSRE.2023.3265304.

Konstantinos Barmpas, Yannis Panagakis, Dimitrios Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A causal perspective in brainwave foundation models. In *Causality and Large Models @NeurIPS 2024*, 2024a.

Konstantinos Barmpas, Yannis Panagakis, Georgios Zoumpourlis, Dimitrios A Adamos, Nikolaos Laskaris, and Stefanos Zafeiriou. A causal perspective on brainwave modeling for brain–computer interfaces. *Journal of Neural Engineering*, 21(3):036001, may 2024b. doi: 10.1088/1741-2552/ad3eb5.

Ali Bashashati, Mehrdad Fatourechi, Rabab K Ward, and Gary E Birch. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural Engineering*, 4(2):R32–R57, mar 2007. doi: 10.1088/1741-2560/4/2/r03.

A. Biasiucci, R. Leeb, I. Iturrate, S. Perdikis, A. Al-Khodairy, T. Corbet, A. Schnider, T. Schmidlin, H. Zhang, M. Bassolino, D. Viceic, P. Vuadens, A.G. Guggisberg, and J. d. R. Millán. Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. *Nat Commun*, 9, 2018. doi: 10.1038/s41467-018-04673-z.

D.C. Castro, I. Walker, and B. Glocker. Causality matters in medical imaging. Commun 11, 3673. 2020. doi: https://doi.org/10.1038/s41467-020-17478-w.

U. Chaudhary, N. Birbaumer, and A. Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nat Rev Neurol*, 12, 2016. doi: 10.1038/nrneurol.2016.113.

Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

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.

Laura Frølich, Irene Winkler, Klaus-Robert Müller, and Wojciech Samek. Investigating effects of different artefact types on motor imagery BCI. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1942–1945. IEEE, 2015.

Todd C. Handy. *Brain signal analysis advances in neuroelectric and neuromagnetic methods. Cambridge, Mass, MIT Press.* 2009.

Kim Hong-Kyung, Williamson John, Lee Min-Ho, Kwon O-Yeon, Lee Seong-Whan, Fazli Siamac, Kim Yong-Jeong, and Lee Young-Eun. Supporting data for "eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy", 2019.

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

Xiao Jiang, Gui-Bin Bian, and Zean Tian. Removal of artifacts from EEG signals: a review. *Sensors*, 19(5):987, 2019.

B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. doi: 10.1109/10.867928.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, July 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c.

T.P. Luu, S. Nakagome, Y. He, and J.L. Contreras-Vidal. Real-time eeg-based brain-computer interface to a virtual avatar enhances cortical involvement in human. *Sci Rep*, 7, 2017. doi: 10.1038/s41598-017-09187-0.

D.J. McFarland, C.W. Anderson, K.-R. Muller, A. Schlogl, and D.J. Krusienski. Bci meeting 2005-workshop on bci signal processing: feature extraction and translation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):135–138, 2006. doi: 10.1109/TNSRE.2006.875637.

Yuri G. Pavlov, Dauren Kasanov, Alexandra I. Kosachenko, and Alexander I. Kotyusov. "eeg, pupillometry, ecg and photoplethysmography, and behavioral data in the digit span task and rest", 2022.

Hans Reichenbach. *The Direction of Time*. Dover Publications, 1956.

Yousef Rezaeitabar and Ugur Halici. A novel deep learning approach for classification of eeg motor imagery signals. *Journal of Neural Engineering*, 14:016003, 02 2017. doi: 10.1088/1741-2560/14/1/016003.

Eduardo Santamaría-Vázquez, Víctor Martínez-Cagigal, Fernando Vaquerizo-Villar, and Roberto Hornero. Eeg-inception: A novel deep convolutional neural network for assistive erp-based brain-computer interfaces. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28 (12):2773–2782, 2020. doi: 10.1109/TNSRE.2020.3048106.

Gerwin Schalk, Dennis J McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R Wolpaw. BCI2000: a general-purpose brain-computer interface (BCI) system. *IEEE Trans. Biomed. Eng.*, 51(6):1034–1043, June 2004.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

Edgar P. Torres, Edgar A. Torres, Myriam Hernández-Álvarez, and Sang Guun Yoo. Eeg-based bci emotion recognition: A survey. *Sensors*, 20(18), 2020. ISSN 1424-8220. doi: 10.3390/s20185083.

Jose Antonio Urigüen and Begoña Garcia-Zapirain. EEG artifact removal—state-of-the-art and guidelines. *Journal of neural engineering*, 12(3):031001, 2015.

Tao Xu, Yun Zhou, Zi Wang, and Yixin Peng. Learning emotions eeg-based recognition and brain activity: A survey study on bci for intelligent tutoring system. *Procedia Computer Science*, 130: 376–382, 2018. ISSN 1877-0509. doi: j.procs.2018.04.056. The 9th International Conference on Ambient Systems, Networks and Technologies (ANT 2018) / The 8th International Conference on Sustainable Energy Information Technology (SEIT-2018) / Affiliated Workshops.

A CAUSAL DISTRIBUTION SHIFTS IN BCIS

According to Barmpas et al. (2024b), there are many shifts associated with the identified causal variables of interest. Table 5 summarizes some of the various shifts in brainwave decoding as well as the causal sub-module they usually affect.

Type	Related with	Causal Shift to	Affects
Stimulus Shift	Experimental Settings	$P(\mathsf{S})$	Involuntarily Engaged and Exogenous
Shifts on Subject's Intention	Experimental Settings	$P(\mathsf{I})$	Voluntarily Engaged and Endogenous
Shifts on Environmental Conditions	Experimental Settings	$P(\mathsf{E})$	Involuntarily Engaged and Endogenous
Acquisition Shift	EEG Data	$P(\mathsf{X}\|\mathsf{Z})$	All Cases
Subject Shift	Subjects	$P(\mathsf{Z}\|\cdot)$	All Cases
Class Imbalance	Associated Labels	$P(\mathsf{Y})$	Exogenous

Table 5: All possible shifts in brainwave decoding for all causal BCI factorization scenarios. Table adapted from Barmpas et al. (2024b).