Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

Foundation Models for Decoding Brain Activity - Benchmarking

Bachelor's Thesis

Josua Bürki

jbuerki@ethz.ch

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:
Ard Kastrati
Prof. Dr. Roger Wattenhofer

April 11, 2025

Acknowledgements

I would like to express my sincere gratitude to Ard Kastrati for his exceptional guidance, his incredibly swift responses, and continuous support throughout this thesis. His dedication and motivation were a constant source of inspiration. I am also thankful to Maxim Huber and Cheng Xuan for their prior contributions, and to Riccardo Rigoni and Jonas Lauer for their valuable exchange and collaboration. Finally, I would like to thank Professor Roger Wattenhofer for his supervision, and DISCO for providing the computational resources essential to this work.

Abstract

This thesis presents the *Unified EEG Benchmark*, a comprehensive framework for evaluating the generalization performance of foundation models in EEG decoding. Spanning 23 publicly available datasets across motor imagery and clinical classification tasks, the benchmark standardizes task definitions and evaluation protocols to enable rigorous cross-subject and cross-dataset comparisons. We assess classical baselines alongside recent foundation models, including BENDR [1], Neuro-GPT [2], and LaBraM [3]. Results indicate that while traditional approaches remain competitive on constrained tasks, foundation models – particularly LaBraM – demonstrate superior performance in generalization and clinical robustness. The benchmark provides a reproducible platform for advancing generalizable EEG decoding and supports the development of clinically relevant brain–computer interfaces.

The code for the benchmark can be found at `https://github.com/b-ausoj/unified_eeg_benchmark`

Contents

Chapter 1

Introduction

The human brain's activity, captured through electroencephalography (EEG), reveals fundamental insights into cognition, neurological disorders, and motor control. As a non-invasive measurement technique, EEG plays a vital role for both clinical applications and brain-computer interfaces (BCIs). However, decoding these neural signals presents a major challenge – EEG patterns vary significantly across individuals, recording sessions, and hardware setups. Traditional machine learning approaches often fail to generalize beyond the specific subjects or conditions they were trained on, limiting their real-world utility.

Recent advances in artificial intelligence have introduced foundation models – large-scale transformer networks pretrained on massive datasets using self-supervised learning, then adapted to specific tasks through fine-tuning. Inspired by their success in domains such as natural language processing, researchers have begun applying these models to EEG analysis, hypothesizing that their ability to learn generalized representations could overcome key challenges like inter-subject variability and data scarcity. Yet, the field lacks standardized benchmarks for rigorously evaluating model performance across subjects and datasets — a critical step toward developing robust and clinically useful EEG decoding systems.

This thesis makes three primary contributions to address this gap.

1. We present a unified EEG benchmark for evaluating foundation models on both motor imagery and clinical classification tasks. The benchmark integrates 23 publicly available datasets, applying standardized but minimal preprocessing and consistent evaluation protocols.
2. We systematically analyze the generalization performance of state-of-the-art foundation models (e.g., BENDR [1], Neuro-GPT [2], LaBraM [3]) across held-out datasets and unseen task configurations.
3. We derive practical insights for improving cross-subject and cross-dataset performance, offering guidance for future work in robust EEG decoding.

Our benchmark builds on the Mother of All BCI Benchmarks (MOABB) [4],

but introduces several key innovations: it focuses specifically on the generalization capabilities of foundation models, extends coverage beyond BCI to include clinical EEG applications, and implements rigorous cross-dataset evaluation protocols – a feature lacking in existing frameworks. The remainder of this thesis is organized as follows: Chapter 2 reviews EEG fundamentals and related work. Chapter 3 describes the design and implementation of the benchmark. Chapter 4 outlines the evaluated models. Chapter 5 presents experimental results, and Chapter 6 concludes with a discussion and directions for future work.

Chapter 2

Background & Related Work

2.1 Overview on Electroencephalography (EEG)

Electroencephalography (EEG) is a typically non-invasive method used to record the brain's electrical activity via electrodes placed along the scalp [5]. This technique captures voltage fluctuations caused by ionic currents in neurons, with signals typically measured in microvolts (μV). Since the electrodes are positioned on the scalp, the recorded signals primarily reflect the activity of superficial cortical neurons, while deeper neurons contribute less. Electrode placement is critical, as different brain regions serve distinct functions. To ensure consistency, electrodes are often arranged according to the International 10-20 system [6], which defines a standardized spatial layout (illustrated in Figure 2.1). Each channel corresponds to the signal from a single electrode. EEG recordings are typically sampled between 200 and 1000 Hz to accurately capture rapid fluctuations in neural activity. Another important consideration is the montage, which defines the referencing strategy. An average reference montage computes the mean potential across all electrodes to reduce local noise, while a linked-ear montage references signals to electrodes placed on the earlobes.

2.2 Real-World EEG Applications

EEG has a wide range of applications, spanning from clinical diagnosis to advanced research in brain–computer interfaces (BCIs) and beyond.

In clinical settings, EEG characteristic spikes and patterns associated with abnormal brain activity, allowing specialists to detect various neurological and psychiatric conditions. For instance, EEG is a vital tool in the diagnosis and management of epilepsy [7], forming the basis for detecting and predicting epileptic seizures [8] and potentially providing early warnings. It is also used to assess brain responses following mild traumatic brain injury (mTBI) [9]. Moreover, waveform shape analysis has emerged as a non-invasive biomarker for monitoring Parkinson's disease [10], and distinct EEG features have been identified as

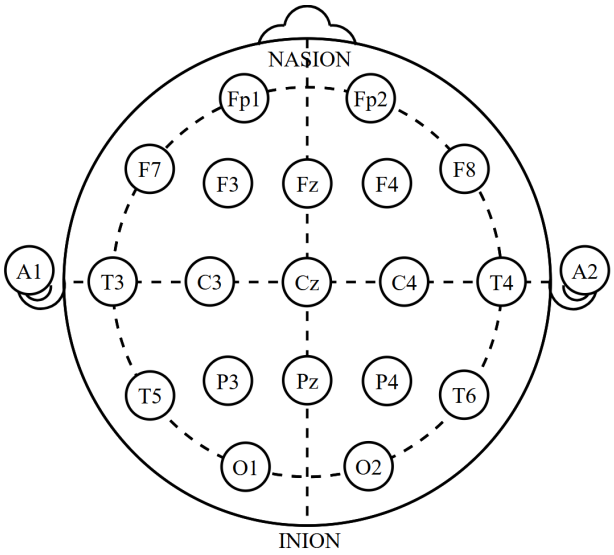

Figure 2.1: Electrode locations of International 10-20 system for EEG recording [5]

potential biomarkers for schizophrenia diagnosis [11].

Beyond clinical applications, EEG is integral to the development of BCI systems, which translate brain activity into commands, thereby establishing a direct communication link between the brain and external devices such as computers, wheelchairs, or robotic exoskeletons. For instance, BCIs have enable users to control cursors [12] and operate electric wheelchairs [13], enhancing autonomy for individuals with limited motor capabilities.

2.3 Challenges in EEG Data Acquisition and Analysis

Despite the wide range of EEG applications in clinical and research contexts, effective use of EEG data remains challenging. In the following section, we outline key obstacles in EEG data acquisition and analysis, including signal variability, low signal-to-noise ratio, limited dataset sizes, hardware discrepancies, and variability in electrode placement.

1. **Inter-Subject Variability** EEG signals are highly individualized, reflecting anatomical, physiological, and functional differences between subjects. Factors such as head shape, scalp thickness, and the exact positioning of electrodes contribute to this variability. As a result, neural patterns recorded during motor imagery tasks can differ significantly between individuals [14].

Moreover, the non-stationarity of EEG signals – with temporal and spectral characteristics varying over time – further complicates the development of models that generalize effectively. These combined factors hinder the creation of classification systems that generalize across subjects (inter-subject) and across sessions within the same subject (intra-subject or cross-session) [15].

2. **Low Signal-to-Noise Ratio** The voltage fluctuations measured in EEG recordings are typically in the microvolt range, making them highly susceptible to interference from various noise sources. Physiological artifacts (e.g., muscle activity, eye blinks) and environmental disturbances further degrade signal quality [16]. This inherently low signal-to-noise ratio poses a major challenge for extracting reliable features and necessitates sophisticated preprocessing and noise reduction techniques.

3. **Data Acquisition and Dataset Size** Data acquisition and dataset size present significant challenges in EEG research. Although the number of publicly available EEG datasets has grown, most still contain recordings from a limited number of subjects – often fewer than ten. Typically, each subject participates in one or two sessions, each comprising only a few runs. Within each run, the experimental task is repeated multiple times, with the number of trials per class usually ranging from 20 to 100. As a result, an average subject may contribute approximately 1,000 trials, each lasting around 3 to 4 seconds.

   This relatively small volume of data is particularly problematic given the high dimensionality of EEG recordings, which usually involve 22 or more channels sampled at 200 Hz or higher. The limited number of trials per subject makes it challenging to obtain statistically robust results, as high-dimensional data requires larger sample sizes for reliable model training and generalization. Similar limitations are observed in clinical EEG studies, where datasets often exhibit small sample sizes, thereby complicating the development models that generalize well across subjects and conditions.

4. **Heterogeneity of Acquisition Hardware** The quality and fidelity of EEG recordings are strongly influenced by the acquisition hardware. Modern EEG systems range from high-end research-grade devices with advanced amplifiers to portable, consumer-grade solutions – each with distinct performance characteristics. One key factor is the type of electrode used: gel-based, water-based, and dry electrodes differ in their electrical properties, which can significantly affect the power spectral density (PSD) of the signals. For example, studies have shown that wet electrodes may result in a greater percentage decrement in PSD following stimulation compared to dry electrodes [17].

5. **Electrode Placement and Montage Variability** In cross-dataset EEG

analyses, variability in electrode placement and referencing montages poses a major challenge. The International 10-20 system is widely recognized as a standard for electrode positioning, typically involving 19 channels along with ground and reference electrodes. However, many studies deviate from this baseline – either by extending to denser configurations such as the 10–10 or 10–5 systems, or by using custom arrangements tailored to specific brain regions. This variation leads to differences in both the number and spatial location of channels, complicating direct comparisons across datasets. Furthermore, some datasets do not provide complete documentation of electrode configurations or channel ordering, introducing ambiguity that not only undermines reproducibility and cross-dataset alignment, but also makes it difficult to incorporate such datasets into standardized benchmarks.

2.4 Advancements in EEG Decoding

EEG decoding has traditionally relied on structured signal processing pipelines that include spatial filtering, feature extraction, and classical machine learning. One of the most influential techniques is Common Spatial Patterns (CSP) [18], which computes spatial filters that maximize the variance difference between two classes – typically used in motor imagery classification. Variants such as Filter Bank CSP (FBCSP) [19] and extensions based on Riemannian geometry have further improved performance by leveraging covariance structures and frequency-specific information [20]. These methods, when combined with simple classifiers like Linear Discriminant Analysis (LDA) or Support Vector Machines (SVM), have shown strong performance, particularly in subject-specific settings. However, their effectiveness often declines in cross-subject or cross-dataset scenarios without extensive tuning or domain adaptation, limiting their applicability in more generalizable EEG decoding pipelines [21].

The emergence of deep learning marked a shift away from handcrafted feature pipelines toward end-to-end learning directly from raw or minimally preprocessed EEG signals. Convolutional Neural Networks (CNNs), in particular, have shown strong performance in decoding tasks by learning spatial and temporal filters jointly. These models have been successfully applied to motor imagery, event-related potentials, and affective state classification [22] [23].

More recently, foundation models – including large-scale neural networks and transformer-based architectures – have begun to reshape EEG decoding. Pretrained on massive and diverse datasets using self-supervised objectives, these models aim to learn general-purpose representations that can be adapted to a variety of tasks and subject populations [1] [2] [3]. Their ability to extract hierarchical, task-agnostic features holds promise for overcoming key challenges such as inter-subject variability and dataset heterogeneity. Nonetheless, foundation

models also bring new challenges, including high computational requirements, the need for large pretraining corpora, and limited interpretability – all of which remain active areas of research.

2.5 Limitations of Existing Benchmarks

Several benchmarking initiatives have played an important role in advancing EEG-based BCIs. One of the most influential is the Mother of All BCI Benchmarks (MOABB) [4], which offers a standardized framework for evaluating BCI algorithms across a wide range of publicly available datasets. MOABB addresses key challenges in reproducibility by unifying preprocessing steps, task definitions, and evaluation metrics within a consistent, open-source ecosystem.

More recently, the Benchmarks for EEG Transfer Learning (BEETL) [24] competition has aimed to foster the development of subject-independent and meta-learning approaches. BEETL focuses on cross-subject generalization and introduces a competitive setting for comparing transfer learning methods across heterogeneous datasets.

Despite these contributions, important limitations remain. MOABB primarily evaluates within individual datasets and sessions, lacking support for cross-dataset generalization – a crucial requirement for real-world deployment. BEETL introduces cross-subject validation but is restricted to a single classification task and does not generalize across EEG paradigms. Critically, both MOABB and BEETL are focused exclusively on BCI use cases and omit clinical applications such as epilepsy, Parkinson's disease, or schizophrenia – areas where EEG has significant diagnostic relevance.

These gaps highlight the need for a unified, extensible benchmark that evaluates models across both BCI and clinical domains, supports diverse tasks and datasets, and explicitly emphasizes generalization to unseen subjects and recording conditions – the central goal of the benchmark introduced in this thesis.

Chapter 3

Benchmark Framework & Methodology

This chapter presents the *Unified EEG Benchmark* a framework designed to systematically evaluate EEG decoding models across both BCI and clinical applications. The goal is to enable transparent, reproducible, and standardized comparisons between traditional and foundation models under real-world conditions – specifically addressing the lack of unified protocols in existing EEG evaluation frameworks. To achieve this, the benchmark integrates a broad range of publicly available datasets, grouped into two main domains: motor imagery classification and clinical diagnostics. It establishes consistent but minimal preprocessing, defines standardized metadata conventions and task structures, and ensures compatibility across heterogeneous datasets. To promote real-world relevance, the framework emphasizes rigorous evaluation of cross-subject and cross-dataset generalization. This chapter describes the benchmark's architecture and implementation, details the included BCI and clinical tasks, and outlines the evaluation protocol, including dataset splitting strategies and performance metrics.

3.1 Benchmark Architecture and Implementation

The benchmark is implemented in Python 3 and structured for modularity, extensibility, and ease of use. It is organized into three main components: `datasets`, `models`, and `tasks`, along with two auxiliary modules: `utils` and `enums`. To accommodate the structural differences between motor imagery (BCI) and clinical EEG datasets (e.g., trial-based vs. continuous recordings), each core directory is further divided into bci and clinical subdirectories. This structure simplifies code navigation and supports consistent integration of new datasets, models, and tasks.

3.1.1 Data Access and Preprocessing

The benchmark provides a unified interface for loading and handling EEG datasets stored in a variety of file formats, including `.mat` and `.edf`. Loaded data is returned either as NumPy arrays or as MNE Raw objects, depending on dataset's structure. For downstream processing, NumPy arrays are standardized to have shape `(n_samples, n_channels, n_sample_length)`.

Each dataset is accompanied by a metadata object containing essential information such as channel names, sampling rate and dataset origin – supporting transparency and reproducibility. All EEG signals are standardized to the microvolt (μV) scale. To reduce memory usage and computational overhead, signals are typically resampled to 250 Hz. To maintain a realistic setting, no additional preprocessing such as filtering or artifact removal is applied. However, for datasets accessed via the MOABB library, a bandpass filter from 8–35 Hz is applied internally. While this narrower band may limit generalizability in some tasks, it was retained to ensure compatibility with the MOABB pipeline and reproducibility of results.

3.1.2 Tasks Definition

Tasks constitute 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.

3.1.3 Model Interface and Training

Considering the diverse range of available EEG classification models and their varying training requirements, the benchmark adopts an abstract interface approach rather than prescribing a fixed training pipeline. Users are expected to implement their models following the provided `AbstractModel` interface, which requires two methods:

- `fit(X, y, meta)`: trains the model on the provided data, labels, and metadata.
- `predict(X, meta)`: generates predictions based on new data and associated metadata.

This flexible design allows researchers to easily integrate novel methods into the benchmark, fostering direct comparison across a wide range of modeling approaches.

3.1.4 Evaluation Scripts

Once predictions are made, evaluation scripts compute the scores on the defined `scikit-learn` metrics for this task, optionally generating visual summaries such as confusion matrices and classification reports to support qualitative analysis and interpretation of model predictions.

3.1.5 User Interaction and Benchmark Execution

The benchmark is designed to be both user-friendly and flexible, enabling straightforward integration and evaluation of EEG models. Users can interact with the framework through two primary modes: a Python-based script interface and a command-line interface (CLI).

In the script-based approach, users define the benchmark configuration by instantiating task and model objects within a dedicated execution script. This configuration specifies the tasks to evaluate, the models to benchmark, and additional parameters such as the random seed for reproducibility. Once defined, the benchmark framework automatically handles the complete pipeline – including dataset download, data loading, model training, prediction, and metric computation. For greater convenience and automation, the CLI enables users to execute experiments by specifying task names, model identifiers, and optional parameters directly via terminal commands.

3.1.6 Reproducibility and Environment

Ensuring reproducible results is a cornerstone of the Unified EEG Benchmark. To this end, the benchmark provides:

- Explicitly predefined train/test splits within each task.
- A fixed random seed applied consistently across Python's `random`, NumPy, and PyTorch libraries to ensure deterministic behavior.
- A Conda environment specification, clearly outlining all dependencies and their versions, simplifying setup and promoting reproducibility.

3.2 Motor Imagery Classification

Motor imagery tasks require participants to imagine performing specific actions – such as moving limbs or facial muscles – without any overt physical execution. During these trials, brain regions involved in motor planning and execution become active, and the resulting EEG signals can be interpreted to infer the

intended movement. The primary objective is to determine whether predictive models can accurately classify these imagined movements from the recorded EEG data. The tasks covered in this benchmark range from simple binary distinctions (e.g., left vs. right hand) to multi-class scenarios (e.g., left hand, right hand, feet, and tongue).

3.2.1 Datasets and Protocol Variability

Table 3.1 provides an overview of the publicly available motor imagery (MI) datasets employed in this benchmark. For each dataset, it lists the MI classes included (e.g., left hand (L), right hand (R), feet (F), tongue (T), fingers (5F), the number of EEG channels recorded, the number of subjects involved in the study, the sampling frequency in Hertz (Hz), the trial length in seconds, and the total number of trials. These datasets are publicly available and can be downloaded via the MOABB interface or directly from their original sources, such as Figshare.

Each dataset differs in the number of channels, sampling frequency, trial length, total number of trials, and the types of MI classes represented. For instance, BCICompIV2a features 22 channels sampled at 250 Hz over 4-second trials, encompassing four MI classes (L, R, F, T). In contrast, BCICompIV2b uses only 3 channels and supports 2 classes (L vs. R), resulting in fewer total trials. Other datasets, like Schirrmeister2017, incorporate as many as 128 channels, while others record fewer than 10. Sampling frequencies span from 160Hz up to 512Hz, and trial durations range from 1 to 7 seconds.

These discrepancies reflect methodological and practical variations across research groups. While such heterogeneity poses challenges – such as harmonizing preprocessing, reconciling channel configurations, and aligning class definitions – it also strengthens the benchmark by introducing a wide range of conditions. Consequently, successful approaches must demonstrate robust generalization across diverse recording setups and participant populations.

3.2.2 Tasks

Left Hand vs. Right Hand.

One of the most extensively studied paradigms in MI research involves classifying EEG signals corresponding to the motor imagery of the left hand versus the right hand. The relative ease of this task arises from the distinctly contralateral cortical activity, particularly under electrodes C3, C4, and Cz in the International 10-20 system. As a result, left- vs. right-hand MI classification is widely regarded as a foundational and deployable paradigm in EEG-based brain-computer interfaces [25] [26] [33]. The task is constructed from nine publicly available BCI datasets (listed in Table 3.2), each with subject-level training and test splits and

Motor Imagery Datasets						
Name	**Classes**	**No. Ch.**	**No. Sub.**	**Freq.** [Hz]	**Trial len.** [s]	**Total Trials**
BCICompIV2a [25]	L, R, F, T	22	9	250	4.0	62208
BCICompIV2b [26]	L, R	3	9	250	4.5	32400
Weibo2014 [27]	L, R, F	60	10	200	4.0	5600
Cho2017 [28]	L, R	64	52	512	3.0	9800
Liu2022 [29]	L, R	29	50	500	4.0	2000
Schirrmeister2017 [30]	L, R, F	128	14	500	4.0	13440
PhysionetMI [31]	L, R, F	64	109	160	3.0	69760
Zhou2016 [32]	L, R, F	14	4	250	5.0	11496
Kaya2018 [33]	L, R, F, T, 5F	21	13	200	1.0	60000
Barachant2012 [34]	R, F	16	8	512	3.0	480
Faller2012 [35]	R, F	13	12	512	5.0	14400
Scherer2015 [36]	R, F	30	9	256	7.0	7200

Table 3.1: Overview of datasets used in Motor Imagery tasks

balanced class distributions. Notably, the Zhou2016 dataset is held out entirely from training to evaluate generalization to unseen recording conditions and subjects.

Task Overview: Left Hand vs. Right Hand MI		
Dataset	**Train Distribution**	**Test Distribution**
BCICompIV2a [25]	1152/1152	144/144
BCICompIV2b [26]	2520/2520	740/740
Weibo2014 [27]	630/630	160/160
Cho2017 [28]	940/940	500/500
Liu2022 [29]	180/180	100/100
Schirrmeister2017 [30]	2610/2612	760/760
PhysionetMI [31]	1975/1940	431/422
Zhou2016 [32]	-	600/599
Kaya2018 [33]	4000/3971	302/340

Table 3.2: Train-test class distribution for the Left Hand vs. Right Hand MI task. The numbers represent the left/right distribution in the training and testing sets.

Right Hand vs. Feet

The second binary task focuses on distinguishing between motor imagery involving the right hand and the feet. This contrast tests a model's ability to separate upper-limb from lower-limb cortical activity, which are characterized by distinct spectral and spatial patterns. As shown in Table 3.3, the task includes data

from nine datasets, with Zhou2016 again used as a held-out dataset for out-of-distribution generalization testing.

Task Overview: Right Hand vs. Feet MI		
Dataset	**Train Distribution**	**Test Distribution**
BCICompIV2a [25]	1008/1008	288/288
Weibo2014 [27]	630/630	160/160
Schirrmeister2017 [30]	2612/2612	760/760
PhysionetMI [31]	1953/1940	429/422
Zhou2016 [32]	-	601/599
Kaya2018 [33]	6509/3321	2501/1272
Barachant2012 [34]	120/120	40/40
Faller2012 [35]	2000/2000	600/600
Scherer2015 [36]	475/475	155/155

Table 3.3: Train-test class distribution for the Right Hand vs. Feet MI task. The numbers represent the right hand and feet distribution in the training and testing sets.

Left Hand vs. Right Hand vs. Feet vs. Tongue

Expanding upon the simpler binary tasks, this multi-class scenario involves four types of motor imagery: left hand, right hand, feet, and tongue. By introducing facial muscle activity (tongue movement), the task requires models to identify more nuanced cortical activation patterns, thereby increasing classification complexity. A particular challenge in this task arises from differences in electrode layouts and reference systems across the two datasets, which can affect spatial signal characteristics and model performance. The task includes data from two datasets, as indicated in Table 3.4.

Task Overview: Left Hand vs. Right Hand vs. Feet vs. Tongue MI		
Dataset	**Train Distribution**	**Test Distribution**
BCICompIV2a [25]	1008/1008/1008/1008	288/288/288/288
Kaya2018 [33]	3295/3354/3321/3514	1269/1277/1272/1333

Table 3.4: Class distributions for the MI task with four classes: Left Hand, Right Hand, Feet, and Tongue. The distributions are given as left/right/feet/tongue counts.

Five Fingers

This task targets the classification of imagined movements of individual fingers: thumb, index, middle, ring, and little. Compared to conventional motor im-

agery tasks, it introduces a higher level of granularity, as models must distinguish between fine motor representations that are anatomically adjacent within the sensorimotor cortex. The close proximity and partial overlap of these cortical activations make feature extraction and classification particularly challenging. This setting serves as a strong test of a model's ability to capture subtle neural differences and generalize across subjects. Unlike the previous ones, this tasks consists of only the Kaya2018 dataset as shown in Table 3.5.

Task Overview: Five Fingers MI		
Dataset	**Train Distribution**	**Test Distribution**
Kaya2018 [33]	2965/3190/3324/3201/3492	343/371/392/377/416

Table 3.5: Class distributions for the MI task with five classes. The distributions are given as thumb/index finger/middle finger/ring finger/little finger counts.

3.3 Clinical Condition Classification

In addition to motor imagery decoding, the benchmark also evaluates model performance on clinical EEG analysis tasks. These tasks involve classifying full-length EEG recordings to detect and differentiate a range of neurological and psychiatric conditions. Unlike typical BCI paradigms, which focus on event-related responses or short trials, the clinical tasks operate on continuous EEG data, better reflecting the nature of real-world diagnostic use cases.

3.3.1 Datasets and Protocol Variability

The benchmark incorporates a diverse set of publicly available clinical datasets spanning several diagnostic paradigms, including abnormal brain activity, epilepsy, Parkinson's disease (PD), Obsessive-Compulsive Disorder (OCD), Mild Traumatic Brain Injury (mTBI), and schizophrenia. Table 3.6 provides a detailed summary of the datasets used. These datasets vary widely in technical specifications – such as the recording durations and clinical annotation standards – as well as in demographic characteristics, including age and gender. The datasets are sourced from the PRED+CT [37] repository and the large Temple University Hospital (TUH) EEG Data Corpus [38].

This heterogeneity presents significant challenges for generalization, particularly in developing models that are robust across acquisition protocols and patient populations. At the same time, it provides a valuable testbed for assessing the real-world applicability of EEG decoding methods in complex and clinically relevant settings.

Clinical Datasets					
Name	**Paradigm**	**No. Ch.**	**No. Sub.**	**Freq.** [Hz]	**Total len.** [h]
Cavanagh2018a [39]	PD	60	50	500	21
Cavanagh2018b [39]	PD	63	54	500	4
Albrecht2019 [40]	Sz	62	78	1000	51
Singh2018 [41]	PD	60	56	500	108
Brown2020 [42]	PD	60	56	500	31
Gründler2009 [43]	OCD	64	46	500	22
Cavanagh2019 [44]	mTBI	60	62	500	57
Singh2020 [45]	PD	63	39	500	7
Singh2021 [46]	PD	63	129	500	58
TUAB [38]	Abnormal	19-23	2383	256	1142
TUEP [47]	Epilepsy	19-23	200	256	625

Table 3.6: Clinical Datasets Summary

3.3.2 Tasks

Abnormal Detection

This task involves binary classification of EEG recordings as either normal or abnormal, based on the TUH Abnormal EEG Corpus (TUAB v3.0.1) [38] – a curated subset of the larger TUH EEG Corpus comprising approximately 60 GB of clinical EEG data from a demographically diverse population. Each recording is annotated by medical experts based on standardized diagnostic criteria, capturing a broad spectrum of neurological abnormalities. While TUAB is widely used for benchmarking model performance [48], prior approaches often relied on classifying short, segmented EEG windows of varying lengths. In contrast, this benchmark evaluates models on full-length recordings, aligning more closely with real-world diagnostic workflows. This avoids ambiguity when abnormality is only partially visible in isolated segments and provides a more stable context for classification. The data split is predefined by the dataset, with 1,371 normal and 1,346 abnormal recordings in the training set, and 150 normal and 126 abnormal recordings in the test set.

Epilepsy Detection

The objective of this task is identifying EEG recordings from patients with epilepsy based on characteristic patterns such as spikes, sharp waves, and abnormal rhythmic activity. It consits of the TUH EEG Epilepsy Corpus (TUEP v2.0.1) [47], a balanced subset of the larger TUH EEG dataset, containing recordings from 100 epileptic and 100 non-epileptic subjects. However the overall

dataset contains significantly more recordings from epileptic patients than from controls (1,651 to 390), presenting a substantial class imbalance challenge. With approximately 72 GB of data and well-defined annotations, TUEP offers a standardized and widely adopted benchmark for evaluating seizure detection models. As with the abnormal EEG task, the evaluation is performed on full recordings rather than short temporal segments. The train/test split is created as a random but balanced subset of subjects, using 75% for training and 25% for testing.

Parkinson's Disease

The Parkinson's disease (PD) classification task aims to distinguish EEG recordings from individuals with PD and healthy controls. Across the datasets used in this benchmark, subjects engaged in a variety of cognitive and motor tasks designed to probe PD-related neural signatures – each dataset adopting a different experimental protocol. A key challenge in this task arises from the heterogeneity in recording conditions, including participation of PD subjects both ON and OFF medication, as well as a demographically diverse population. These factors introduce substantial variability into the EEG signals, making this task a demanding test of a model's ability to extract disease-relevant patterns that generalize across contexts. The class distributions are shown in 3.7 with the Singh2020 dataset held out entirely to evaluate cross-dataset generalization.

Task Overview: Parkinon's Disease (PD) Clinical		
Dataset	**Train Distribution**	**Test Distribution**
Cavanagh2018a [39]	33/20	16/8
Cavanagh2018b [39]	20/20	8/8
Singh2018 [41]	20/19	8/8
Brown2020 [42]	40/20	16/8
Singh2020 [45]	-	26/13
Singh2021 [46]	55/23	24/10

Table 3.7: Class distributions for the PD Clinical task. The distributions are given as counts for each class (e.g., PD/control) for the training and testing sets.

Obsessive-Compulsive Disorder (OCD)

This task involves classifying EEG recordings from individuals with obsessive-compulsive disorder (OCD) versus healthy controls. OCD is associated with hyperactivity in cortico-striatal circuits, particularly the anterior cingulate cortex, which contributes to amplified error signals and altered reinforcement learning dynamics. These mechanisms manifest as larger error-related negativities (ERNs) in EEG, though the patterns are often spatially diffuse [43]. Combined with the small sample size – a single dataset containing recordings from 24 individuals

with OCD and 22 controls – these characteristics make OCD classification a challenging task, requiring models that can detect fine-grained, distributed neural differences under data-scarce conditions. The train and test class distributions are 18/17 and 6/5 (OCD/control), respectively.

Mild Traumatic Brain Injury (mTBI)

This task involves classifying EEG recordings from individuals with a history of mild traumatic brain injury (mTBI) and healthy controls. mTBI can lead to lasting disruptions in executive functioning, which may be reflected in altered event-related potentials [44]. However, these neural changes are typically subtle and vary across individuals, making them difficult to detect reliably. Combined with a small sample size – a single dataset containing 73 recordings labeled as mTBI and 104 as control – this task poses a hurdle for models to extract clinically meaningful features under conditions of high signal variability and noise. The train and test class distributions are 64/85 and 9/19 (mTBI/control), respectively.

Schizophrenia

This task aims to classify EEG recordings from individuals with schizophrenia and healthy controls. While schizophrenia is associated with impairments in reward processing and effort-based decision-making, these effects are often subtle at the neural level. EEG signatures such as mid-frontal theta responses may reflect increased cognitive effort, but often lack strong differentiability across subjects [40]. As a result, this task presents a high-level challenge for models attempting to detect nuanced cognitive dysfunction in a low-sample regime. The dataset used for this task comprises 76 recordings in total, with 25 schizophrenia and 38 control recordings used for training, and 6 and 7, respectively, in the test set.

3.4 Evaluation Protocol

To ensure rigorous and reproducible evaluation across a wide range of EEG decoding tasks, we adopt a standardized evaluation protocol based on fixed data splits. For each dataset in the benchmark, non-overlapping training and testing sets are defined at the subject level, ensuring that no individual appears in both subsets. Special attention is given to individuals who appear in multiple datasets. In such cases, each subject is assigned exclusively to either the training or testing partition across all datasets. This design is crucial for assessing a model's ability to generalize beyond seen individuals and prevent any data leakage that could inflate performance.

Where applicable, we perform held-out dataset evaluations. In these scenarios, an entire dataset is excluded from all training and validation steps and used solely for final evaluation. This setting simulates a real-world deployment scenario, where a model must be applied to an entirely new recording protocol, subject population, or clinical cohort without prior exposure.

Model performance is evaluated using a comprehensive suite of metrics: Balanced Accuracy, Precision, Recall, Weighted F1 Score, ROC AUC, and Average Precision (AP). All metrics are computed with attention to class imbalance to ensure robust assessment across both binary and multi-class classification tasks. Throughout this thesis, Balanced Accuracy (in percentage) and Weighted F1 Score (scalar) are reported as the primary metrics. These offer a strong trade-off between sensitivity to class distribution and ease of interpretability for comparative analysis.

Balanced Accuracy

Balanced Accuracy is the average of recall values (true positive rates) across all classes. It is defined as:

$$\text{Balanced Accuracy} = \frac{1}{C} \sum_{i=1}^{C} \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{3.1}$$

where C is the number of classes, TP_i is the number of true positives for class i, and FN_i is the number of false negatives for class i.

Weighted F1 Score

The F1 score for class i is the harmonic mean of precision and recall:

$$\text{F1}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{3.2}$$

where:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \quad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

The weighted F1 score averages the per-class F1 scores, weighted by the number of true instances in each class:

$$\text{Weighted F1} = \sum_{i=1}^{C} \frac{n_i}{N} \cdot \text{F1}_i \tag{3.3}$$

where n_i is the number of true instances in class i, and $N = \sum_{i=1}^{C} n_i$ is the total number of instances across all classes.

Chapter 4

Evaluated Models

In this chapter, we introduce and detail the models implemented and evaluated on the Unified EEG Benchmark. Our approach includes feature extraction methods combined with standard machine learning baselines, as well as recently released foundation models that have been designed to enhance generalization across diverse EEG datasets and tasks.

4.1 Baseline Models

4.1.1 Preprocessing with Common Spatial Patterns (CSP)

Common Spatial Patterns (CSP) is a well-established algorithm for spatial filtering that is widely used in brain–computer interface (BCI) research. CSP learns spatial filters that maximize the variance differences between classes, thereby enhancing class separability. As demonstrated in [18] and reviewed in [21], CSP has proven effective in extracting features that reflect motor imagery-related neural dynamics.

Formally, CSP involves the simultaneous diagonalization of class-specific covariance matrices, represented as follows: Given two classes with covariance matrices $\Sigma^{(1)}$ and $\Sigma^{(2)}$, CSP seeks spatial filters $\mathbf{W}$ that solve the generalized eigenvalue problem:

$$\Sigma^{(1)}\mathbf{W} = \Lambda\Sigma^{(2)}\mathbf{W}$$

where Λ denotes a diagonal matrix of eigenvalues, and columns of $\mathbf{W}$ correspond to the spatial filters. By projecting EEG signals onto these filters, CSP yields component which have maximum variance for one class and minimum for the other.

4.1.2 Preprocessing with the Brainfeatures Toolbox

For clinical analyses, we employed the Brainfeatures toolbox, an open-source resource tailored for comprehensive EEG feature extraction. This toolbox facili-

tates generating a diverse array of features that encapsulate the intricate dynamics of EEG signals. By employing transformations such as the continuous wavelet transform (CWT), discrete wavelet transform (DWT), and discrete Fourier transform (DFT), Brainfeatures captures both time-domain and frequency-domain characteristics of the EEG data.

The extracted features include spectral power across standard frequency bands, including delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–100 Hz), which are indicative of different neural oscillatory activities. Additionally, the toolbox computes statistical measures such as mean, variance, and higher-order moments, reflecting the amplitude distribution and variability of the EEG signals. Temporal features like signal entropy and fractal dimensions provide insights into the complexity and predictability of neural patterns. Thanks to Brainfeatures built-in epoching and feature aggregation mechanisms, it can process EEG recordings of arbitrary duration and produce consistent feature representations across sessions of varying length – making it well-suited for clinical applications.

4.1.3 Support Vector Machines (SVM)

Support Vector Machines (SVM) are employed as a baseline classifier for our EEG tasks. SVMs are well-suited to handle high-dimensional data and to construct decision boundaries that maximize the margin between classes. The use of the linear kernel restricts the model to linear decision boundaries, providing a strong and interpretable baseline while reducing the risk of overfitting in low-data regimes. In our experiments, SVMs serve as a reference point to evaluate the performance of both standard and advanced models on the benchmark tasks.

4.1.4 Linear Discriminant Analysis (LDA)

Linear Linear Discriminant Analysis (LDA) is a classical method in machine learning, known for its simplicity and effectiveness in low-data regimes. LDA computes a linear projection that maximizes class separability by minimizing within-class variance and maximizing between-class variance. It assumes normally distributed features with class-independent covariance matrices – an assumption that often holds approximately for EEG-derived features. In this benchmark, LDA serves to assess the discriminative power of the extracted features and complements the SVM baseline.

4.2 Foundation Models

In addition to conventional baselines, our benchmark incorporates recently proposed foundation models that are inspired by advances in large-scale neural network architectures. These models are typically based on transformer architectures and pre-trained on vast EEG datasets using self-supervised objectives. The goal is to learn general-purpose, high-dimensional representations that can be fine-tuned for a wide range of downstream tasks. Their inherent ability to capture complex patterns and adapt across heterogeneous datasets has potential to enhance generalization on both motor imagery and clinical tasks. The concept of foundation models originates from the success of Large Language Models (LLMs), and similar architectural principles are now being applied to neurophysiological data. Evaluating these models against traditional baselines will provide valuable insights into their strengths and limitations.

4.2.1 BENDR

BErt-like Neurophysiological Data Representations (BENDR), proposed by Kostas et al. in 2021 [1], adapts language modeling techniques and architectures. It was one of the first to implement self-supervised training styles from automatic speech recognition to pre-train on large-scale, unlabeled EEG datasets.

The architecture of BENDR closely follows that of `wav2wec 2.0` [49] and is comprised of two stages: a feature extraction stage and a transformer-based sequence modeling stage. In the first stage, raw EEG signals are drastically downsampled into a new sequence of vectors using a stack of short-receptive-field 1D convolutional layers, creating compact representations called BENDR embeddings. These embeddings capture the temporal and spatial characteristics of the EEG signals. The second stage uses a transformer encoder to model the relationships between these embeddings, using a contrastive learning objective. Specifically, the model is trained to predict masked EEG segments by contrasting the correct embeddings with distractors, allowing it to learn meaningful representations of brain activity.

BENDR was pre-trained on the Temple University Hospital (TUH) EEG corpus [38], a large-scale, publicly available dataset containing over 10,000 subjects. The model was then fine-tuned on several downstream tasks, including motor imagery classification and sleep stage classification. BENDR yielded accurate and generalizable representations, outperforming previous models. However BENDR has some limitations as it only focuses on a fixed 19 EEG channels and a fixed 60-second sample length. In our experiments, we use the officially released pre-trained BENDR model and apply the published weights provided by the authors.

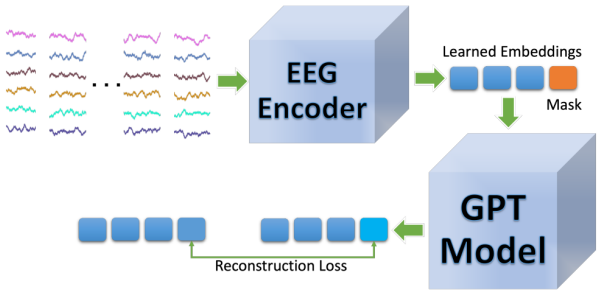

Figure 4.1: Architecture of the Neuro-GPT model [2]

4.2.2 Neuro-GPT

The Neuro-GPT model, proposed by Cui et al. in 2024 [2], is a foundation model designed to address the challenges of data scarcity and heterogeneity in EEG for BCI tasks.

The architecture of Neuro-GPT consists of an EEG encoder and a GPT model. Raw EEG data is first divided into fixed-length chunks, which are treated as tokens. The EEG encoder, composed of convolutional and transformer layers, extracts spatio-temporal features from the raw EEG chunks, generating lower-dimensional embeddings. These embeddings serve as denoised representations of the EEG signals and are fed into the GPT model. The GPT model then learns to predict masked tokens, capturing meaningful temporal correlations and patterns in the data.

After pre-training on the TUH EEG corpus, the model was fine-tuned on a motor imagery classification task with data from nine subjects. Experiments demonstrating that the pre-trained model significantly outperformed models trained from scratch, showcasing that large-scale pre-training enhances performance in downstream tasks by capturing fundamental EEG features that generalize across datasets. For our benchmark, we use the officially released pre-trained weights and evaluate the encoder-only variant of Neuro-GPT, as it achieved the best results in the original paper.

4.2.3 LaBraM

The Large Brain Model (LaBraM), proposed by Jiang et al. in 2024 [3], is a unified foundation model for EEG. LaBraM enables cross-dataset learning by segmenting EEG signals into channel patches. It employs Vector-quantized neural spectrum prediction to train a neural tokenizer, which encodes continuous raw EEG patches into compact neural codes. During pre-training, the neural Transformers are trained to predict the original neural codes for masked EEG channel patches.

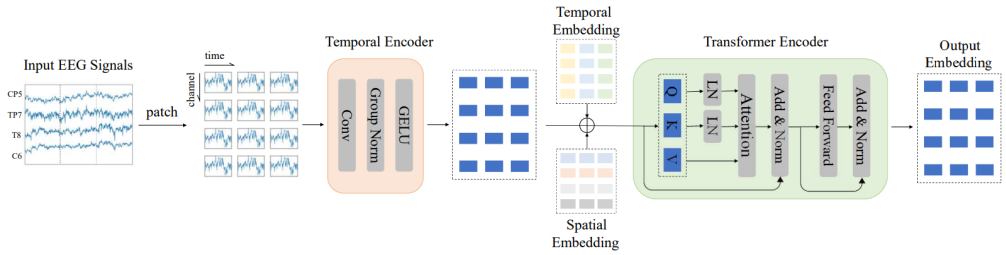

Figure 4.2: Architecture of LaBraM [3]

The architecture of LaBraM is based on a neural Transformer, which processes EEG signals in a structured manner. All input EEG signals are first segmented into patches through a fixed-length time window. Then a temporal encoder is applied to each patch to extract temporal features. Afterwards, temporal and spatial embeddings are added to the patch features to carry temporal and spatial information. In the final step, the sequence of embeddings is passed into the Transformer encoder by patch-wise attention to obtain the output embeddings. One key strength of LaBraM is its ability to handle varying EEG channel configurations.

To evaluate the scalability of LaBraM, the authors created three model variants with varying parameter sizes: LaBraM-Base (5.8M parameters), LaBraM-Large (46M parameters), and LaBraM-Huge (369M parameters). These models are pre-trained on about 2,500 hours of various EEG signals from 20 different datasets. Following pre-training, the models were fine-tuned on four downstream tasks such as abnormal detection, event classification, emotion recognition, and gait prediction, achieving state-of-the-art performance across tasks. In our experiments, we use the pre-trained LaBraM-Base model with the officially released weights, as it is the only variant made publicly available by the authors.

Chapter 5

Results

5.1 Motor Imagery Classification Results

5.1.1 Left Hand vs. Right Hand

The first task evaluates the ability of EEG decoding models to distinguish between imagined left- and right-hand movements. Table 5.1 presents the performance of baseline and foundation models when trained on the combined set of eight datasets and evaluated across all test sets. Overall, performance is relatively consistent, with balanced accuracies ranging from 64.9% to 67.2%, and weighted F1 scores showing a similar trend. LaBraM achieved the highest balanced accuracy at 67.2% followed by BENDR with 66.5%. Interestingly, CSP combined with SVM or LDA performed comparably to the more complex foundation models. This may be attributed to the fact that left- vs. right-hand motor imagery elicits spatially distinct cortical activations, making the task relatively easier to classify. However, no model exceeded 70% accuracy, suggesting substantial variability across the different datasets.

Generalization performance on the held-out Zhou2016 dataset is reported in Table 5.2. In this setting, SVM achieved the highest balanced accuracy (78.5%) and weighted F1 score (0.783), outperforming all other models. Among the foundation models, LaBraM performed competitively with 73.5%, while Neuro-GPT underperformed substantially, with a balanced accuracy of 51.8% and a notably low weighted F1 score of 0.392. The relatively higher performance across all models on Zhou2016 may be due to the fact that subjects in this dataset were explicitly trained on motor imagery tasks, resulting in clearer and more consistent EEG patterns.

Performance Across All Datasets		
Models	**Balanced Accuracy**	**Weighted F1**
CSP+SVM	0.6647	0.6647
CSP+LDA	0.6610	0.6598
BENDR	0.6651 ± 0.0107	0.6644 ± 0.0109
NeuroGPT	0.6487 ± 0.0051	0.6474 ± 0.0050
LaBraM	**0.6722** ± 0.0066	**0.6713** ± 0.0067

Table 5.1: Model performances on Left Hand vs. Right Hand MI task across all datasets

Performance on Zhou2016		
Models	**Balanced Accuracy**	**Weighted F1**
CSP+SVM	**0.7849**	**0.7827**
CSP+LDA	0.7616	0.7565
BENDR	0.7218 ± 0.0351	0.7213 ± 0.0358
NeuroGPT	0.5182 ± 0.0208	0.3916 ± 0.0769
LaBraM	0.7350 ± 0.0285	0.7320 ± 0.0329

Table 5.2: Model performances on Left Hand vs. Right Hand MI task on the held-out dataset Zhou2016

5.1.2 Right Hand vs. Feet

The Right Hand vs. Feet MI task tests the models ability to separate upper-limb from lower-limb cortical activity. Performance across the test set is summarized in Table 5.3. The classical models CSP+SVM and CSP+LDA performed poorly, with balanced accuracies below 60%, likely due to the limited overlap in electrode configurations across datasets and the resulting low spatial resolution. In contrast, LaBraM and BENDR again achieved the highest performance, with balanced accuracies of 73.8% and 74.6%, respectively, and weighted F1 scores above 0.74. Neuro-GPT reached 64.4%, outperforming the classical baselines but lagging behind the other foundation models.

Table 5.4 reports model performance on the held-out Zhou2016 dataset. BENDR achieved the highest balanced accuracy at 74.5% and the highest weighted F1 score at 0.741, maintaining strong performance despite the distribution shift. LaBraM and CSP+LDA also performed well, with balanced accuracies of 71.8% and 71.4%, respectively. In contrast, CSP+SVM and Neuro-GPT showed a marked performance drop, both falling below 51% accuracy and exhibiting poor class separation.

Performance Across All Datasets		
Models	**Balanced Accuracy**	**Weighted F1**
CSP+SVM	0.5800	0.5642
CSP+LDA	0.5682	0.5749
BENDR	**0.7462** ± 0.0035	**0.7505** ± 0.0032
NeuroGPT	0.6440 ± 0.0000	0.6494 ± 0.0000
LaBraM	0.7383 ± 0.0069	0.7426 ± 0.0073

Table 5.3: Model performances on Right Hand vs. Feet MI task across all datasets

Performance on Zhou2016		
Models	**Balanced Accuracy**	**Weighted F1**
CSP+SVM	0.5058	0.3452
CSP+LDA	0.7140	0.7101
BENDR	**0.7446** ± 0.0105	**0.7409** ± 0.0115
NeuroGPT	0.5075 ± 0.0000	0.3535 ± 0.0000
LaBraM	0.7175 ± 0.0144	0.7053 ± 0.0187

Table 5.4: Model performances on Right Hand vs. Feet MI task on the held-out dataset Zhou2016

5.1.3 Left Hand vs. Right Hand vs. Feet vs. Tongue

The four-class motor imagery task introduces a higher level of complexity, requiring models to distinguish between left hand, right hand, feet, and tongue. As shown in Table 5.5, all models exhibited a performance drop relative to the binary tasks. Classical models such as CSP+SVM and CSP+LDA achieved balanced accuracies not far from chance level (28–29%) and weighted F1 scores below 0.28. Foundation models performed notably better, with LaBraM achieving the highest balanced accuracy and F1 score (63.8% and 0.639 respectively), followed by BENDR with 62.5% and 0.623.

The authors of Neuro-GPT reported strong performance on the BCICompIV2a dataset when fine-tuned on that specific dataset, achieving 64.5% accuracy [2]. However, in our benchmark setting, which includes data from both BCICompIV2a and Kaya2018, Neuro-GPT struggled to generalize and achieved a balanced accuracy of 37.8% and a weighted F1 score of 0.377. This aligns with the trend observed in previous tasks, suggesting that Neuro-GPT lacks the flexibility and representational capacity required to generalize across diverse datasets.

Performance Across Both Datasets		
Models	**Balanced Accuracy**	**Weighted F1**
CSP+SVM	0.2869	0.2592
CSP+LDA	0.2909	0.2756
BENDR	0.6249 ± 0.0029	0.6234 ± 0.0027
NeuroGPT	0.3782 ± 0.0000	0.3769 ± 0.0000
LaBraM	**0.6382** ± 0.0021	**0.6381** ± 0.0022

Table 5.5: Model performances on Left Hand vs. Right Hand vs. Feet vs. Tongue MI task across both datasets

5.1.4 Five Fingers

The five-class finger movement task requires classification of imagined movements of the thumb, index, middle, ring, and little fingers. Table 5.6 shows that classical models such as CSP+SVM and CSP+LDA performed poorly, achieving balanced accuracies near 20%, equivalent to random guessing, and F1 scores below 0.15. This result is expected given the fine spatial resolution required to distinguish cortical activations for adjacent fingers.

Among foundation models, LaBraM achieved the highest balanced accuracy at 35.4% and a weighted F1 score of 0.353, followed closely by BENDR (33.9%, 0.332). Neuro-GPT achieved 23.0% accuracy and an F1 score of 0.196, performing only slightly above chance. This may reflect a lack of fine-tuned spatial sensitivity, particularly since such subtle within-limb distinctions were likely absent from its pretraining corpus.

These findings are consistent with prior work. In the original dataset publication by Kaya et al. [33], even intra-subject classification with subject-specific SVMs achieved an average accuracy of only 43%. This aligns with our results and reinforces the conclusion that decoding imagined movements of individual fingers is an inherently low-SNR, high-overlap problem – even under ideal, subject-specific conditions. The poor generalization performance of the foundation models may therefore reflect the fundamental challenge of this paradigm rather than model limitations alone.

Performance on Five Fingers MI		
Models	**Balanced Accuracy**	**Weighted F1**
CSP+SVM	0.2056	0.1312
CSP+LDA	0.1957	0.1455
BENDR	0.3395 ± 0.0082	0.3320 ± 0.0096
NeuroGPT	0.2301 ± 0.0037	0.1956 ± 0.0073
LaBraM	**0.3541** ± 0.0072	**0.3525** ± 0.0071

Table 5.6: Model performances on Thumb vs. Index Finger vs. Middle Finger vs. Ring Finger vs. Little Finger MI task

5.2 Clinical Condition Classification Results

5.2.1 Abnormal Detection

The abnormal EEG classification task involves distinguishing between clinically annotated normal and abnormal recordings from the TUH EEG Abnormal Corpus (TUAB). The results, shown in Table 5.7, indicate that LaBraM achieved the strongest performance, with a balanced accuracy of 83.8% and a weighted F1 score of 0.842. These results exceed those reported in the original LaBraM paper (81.4% [3]), which may be attributed to our evaluation strategy. BENDR, Neuro-GPT, and the classical models based on extracted features from Brainfeatures (explained in 4.1.2) combined with SVM or LDA all showed comparable performance, with balanced accuracies in the range of 67–72%. While none matched LaBraM's results, these models still achieved reasonable performance on full-length clinical EEG recordings.

Performance on TUAB		
Models	**Bal. Acc.**	**Weighted F1**
BF+SVM	0.7222	0.7199
BF+LDA	0.6768	0.6804
BENDR	0.7174 ± 0.0030	0.7224 ± 0.0030
NeuroGPT	0.6956 ± 0.0052	0.6994 ± 0.0052
LaBraM	**0.8376** ± 0.0114	**0.8418** ± 0.0116

Table 5.7: Performance of all models on the abnormal EEG classification task (TUAB). BF refers to features extracted using the Brainfeatures toolbox, as described in Section 4.1.2.

5.2.2 Epilepsy Detection

In this task, models are evaluated on their ability to classify whether an EEG recording corresponds to a patient with epilepsy. As the number of recordings per class is very imbalanced, class weights were applied during training [50]. Table 5.8 summarizes the performance of all models. BENDR achieved the best overall performance, with a balanced accuracy of 74.0% and a weighted F1 score of 0.709. Neuro-GPT followed closely, with similar results. Interestingly, LaBraM underperformed on this task compared to its strong performance on abnormal detection, reaching only 56.5% balanced accuracy. This may indicate a higher sensitivity to class imbalance or a tendency toward overfitting in this setting. Classical models such as SVM and LDA failed to surpass chance-level balanced accuracy (both around 53.1%), although their weighted F1 scores were higher due to bias toward the majority class.

Performance on TUEP		
Models	**Bal. Acc.**	**Weighted F1**
BF+SVM	0.5308	0.6126
BF+LDA	0.5309	0.5934
BENDR	**0.7395** ± 0.0146	**0.7092** ± 0.0288
NeuroGPT	0.7341 ± 0.0099	0.6974 ± 0.0157
LaBraM	0.5649 ± 0.0166	0.6471 ± 0.0184

Table 5.8: Model performances on the Epilepsy Clinical task (TUEP). BF refers to features extracted using the Brainfeatures toolbox, as described in Section 4.1.2.

5.2.3 Parkinon's Disease

This task focuses on detecting Parkinson's Disease (PD) from several different datasets and settings (described in 3.3.2). Table 5.9 reports model performance across the test set. LaBraM achieved the best results, with a balanced accuracy of 66.2% and a weighted F1 score of 0.671, followed closely by Neuro-GPT and the classical baselines. Interestingly, SVM and LDA reached competitive performance levels, with SVM obtaining 63.4% balanced accuracy and 0.657 weighted F1, surpassing BENDR on this task (56.1% and 0.567).

The generalization results on the held-out Singh2020 dataset are presented in Table 5.10. LaBraM again achieved the highest performance, with a balanced accuracy of 73.9% and a weighted F1 score of 0.788, demonstrating strong cross-dataset robustness. Neuro-GPT also performed well (66.2% / 0.712), showing better generalization than in several other clinical tasks. In contrast, BENDR underperformed on this dataset, with both metrics falling below those of the classical baselines.

Performance Across All Datasets		
Models	**Balanced Accuracy**	**Weighted F1**
BF+SVM	0.6339	0.6566
BF+LDA	0.6117	0.6412
BENDR	0.5614 ± 0.0244	0.5670 ± 0.0531
NeuroGPT	0.6371 ± 0.0236	0.6568 ± 0.0149
LaBraM	**0.6622** ± 0.0269	**0.6707** ± 0.0217

Table 5.9: Model performances on PD Clinical task across all datasets. BF refers to features extracted using the Brainfeatures toolbox, as described in Section 4.1.2.

Performance on Singh2020		
Models	**Balanced Accuracy**	**Weighted F1**
BF+SVM	0.5962	0.6540
BF+LDA	0.6154	0.6753
BENDR	0.5885 ± 0.0701	0.6084 ± 0.0808
NeuroGPT	0.6615 ± 0.0375	0.7120 ± 0.0269
LaBraM	**0.7385** ± 0.0349	**0.7878** ± 0.0248

Table 5.10: Model performances on PD Clinical task on the held-out dataset Singh2020. BF refers to features extracted using the Brainfeatures toolbox, as described in Section 4.1.2.

5.2.4 Obsessive-Compulsive Disorder (OCD)

This task evaluates model performance in classifying EEG recordings from individuals with or without Obsessive-Compulsive Disorder (OCD). The results, presented in Table 5.11, show that LaBraM achieved the best overall performance, with a balanced accuracy of 74.0% and a weighted F1 score of 0.743. Neuro-GPT followed closely, demonstrating strong performance despite the small dataset. Interestingly, LDA also performed competitively, achieving a balanced accuracy of 71.7% – outperforming BENDR and SVM, which both underperformed relative to the deep models in this setting.

However, due to the small sample size and high variance across recordings, these results should be interpreted with caution. The limited number of test subjects increases the risk of overfitting and inflates sensitivity to individual variability, reducing the statistical robustness of the observed performance differences.

Performance on OCD Clinical		
Models	**Bal. Acc.**	**Weighted F1**
BF+SVM	0.6333	0.6364
BF+LDA	0.7167	0.7226
BENDR	0.5133 ± 0.0506	0.3717 ± 0.1266
NeuroGPT	0.7033 ± 0.0820	0.6813 ± 0.0890
LaBraM	**0.7400** ± 0.0435	**0.7427** ± 0.0423

Table 5.11: Model performances on OCD Clinical task. BF refers to features extracted using the Brainfeatures toolbox, as described in Section 4.1.2.

5.2.5 Mild Traumatic Brain Injury (mTBI)

This task aims to classify EEG recordings from individuals with or without a history of mild traumatic brain injury (mTBI). As shown in Table 5.12, LDA achieved the best overall performance, with a balanced accuracy of 81.3% and a weighted F1 score of 0.793. This result may suggest that, in low-data scenarios, simpler models with fewer parameters – such as LDA – are less prone to overfitting and better suited to limited training data. In contrast, larger foundation models may require more data to fully leverage their representational capacity. Among the foundation models, LaBraM showed the most consistent performance (balanced accuracy: 65.8%, weighted F1: 0.662), followed by Neuro-GPT. BENDR again underperformed, which may reflect sensitivity to data scarcity or difficulty adapting to the subtle EEG differences associated with mTBI. However, due to the small test set and limited statistical power, these results should be interpreted with caution and considered indicative rather than definitive.

Performance on mTBI Clinical		
Models	**Bal. Acc.**	**Weighted F1**
BF+SVM	0.6228	0.6201
BF+LDA	**0.8129**	**0.7926**
BENDR	0.4906 ± 0.0418	0.4285 ± 0.1976
NeuroGPT	0.5848 ± 0.0709	0.6369 ± 0.0645
LaBraM	0.6579 ± 0.0637	0.6617 ± 0.0712

Table 5.12: Performance on mTBI Clinical task. BF refers to features extracted using the Brainfeatures toolbox, as described in Section 4.1.2.

5.2.6 Schizophrenia

The schizophrenia classification task focuses on distinguishing EEG recordings from individuals diagnosed with schizophrenia and healthy controls. Results are reported in Table 5.13. SVM achieved the highest performance, with a balanced accuracy of 67.9% and a weighted F1 score of 0.681. All three foundation models – BENDR, LaBraM, and Neuro-GPT – underperformed substantially, with none exceeding a balanced accuracy of 55%. BENDR, in particular, performed well below chance level and exhibited high variance across runs, suggesting instability and overfitting.

These results should be interpreted with caution. While schizophrenia is associated with impairments in reward processing and effort-based decision-making, the corresponding EEG markers – such as mid-frontal theta activity – are often subtle and spatially diffuse. Combined with the limited number of subjects and recordings in this dataset, these factors may have made it difficult for foundation models to extract reliable features or learn generalizable patterns. It remains unclear whether the poor performance reflects limitations of the models, the dataset, or the inherent difficulty of the task itself.

Performance on Schizophrenia Clinical		
Models	**Bal. Acc.**	**Weighted F1**
BF+SVM	**0.6786**	**0.6808**
BF+LDA	0.5476	0.5330
BENDR	0.4714 ± 0.0549	0.4214 ± 0.0844
NeuroGPT	0.5452 ± 0.0416	0.5435 ± 0.0409
LaBraM	0.5429 ± 0.0450	0.4632 ± 0.1277

Table 5.13: Performance on Schizophrenia Clinical task. BF refers to features extracted using the Brainfeatures toolbox, as described in Section 4.1.2.

Chapter 6

Conclusion

6.1 Discussion

In this thesis, we have introduced the *Unified EEG Benchmark* for evaluating EEG decoding models across both motor imagery and clinical classification tasks. The benchmark was designed to address limitations in existing frameworks, including fragmented task definitions, inconsistent preprocessing, and a lack of support for evaluating generalization across datasets. The proposed benchmark integrates 23 publicly available datasets, provides standardized tasks and enforces subject-level splits to avoid data leakage.

Across motor imagery tasks, the foundation models LaBraM and BENDR generally outperformed classical baselines, particularly in multi-class settings. For instance, in the four-class finger classification task (Table 5.5), classical models performed near chance level, while LaBraM reached a balanced accuracy of 63.8%, highlighting its relative strength in more complex decoding problems. In contrast, for the simpler binary tasks left- vs. right-hand motor imagery (Table 5.1), classical models achieved competitive results. This suggests that when task structure closely aligns with known cortical activation patterns, simpler approaches can perform comparably to more complex ones. Nevertheless, the overall accuracies – all below 70% – indicate substantial variability across datasets, which none of the models, including foundation models, fully managed to overcome.

In the clinical domain, LaBraM again demonstrated the most consistent performance, outperforming all other models in abnormal EEG detection, OCD classification, and Parkinson's disease generalization. However, classical models performed surprisingly well on several small or imbalanced datasets. For example, LDA achieved the highest balanced accuracy in mTBI detection, and SVM outperformed foundation models in schizophrenia classification. These results suggest that while foundation models can learn richer and more flexible representations, they may be more prone to overfitting in low-resource settings, underscoring the need for regularization and careful validation in clinical applications.

The observed performance differences can be linked to architectural choices and pretraining strategies. LaBraM's patch-wise encoding with spatial and temporal embeddings allowed it to handle varying electrode configurations and likely improved its ability to capture localized brain activity. BENDR also performed well, benefiting from contrastive self-supervised learning, but limiting its effectiveness in heterogeneous datasets due to the fixed channels. Neuro-GPT, on the other hand, underperformed constantly. Its chunk-based tokenization and autoregressive training may be well-suited to language modeling but less effective for EEG, where relevant patterns are more local and less sequential.

Despite these insights, the benchmark has several limitations. First, some clinical tasks – such as OCD, mTBI and schizophrenia detection – are based on small datasets with limited subject diversity, which reduces statistical power and increases performance variance. In these cases, fixed train-test splits may not provide a robust estimate of generalization; future versions of the benchmark should incorporate cross-validation to better account for variability in small datasets. Second, while full-recording evaluation aligns with clinical workflows, it may lead to relevant fine-grained temporal features being overshadowed by broader signal trends. Third, the benchmark currently covers only two EEG paradigms: motor imagery and clinical condition classification. Other application domains – such as emotion recognition or sleep staging – remain unexplored. Fourth, all experiments rely on publicly available datasets, which, while essential for reproducibility, may not fully reflect the heterogeneity found in clinical practice. Finally, training and fine-tuning large foundation models remains computationally demanding, posing challenges for deployment in resource-constrained or real-time settings.

Overall, the benchmark presented in this thesis offers a comprehensive foundation for standardized evaluation in EEG decoding and reveals important trends in model performance, robustness, and limitations. It is intended as a tool to accelerate reproducible research and drive the development of models that are both accurate and clinically meaningful.

6.2 Future Work

Future work should aim to expand the benchmark to include additional EEG paradigms beyond motor imagery and clinical diagnostics. Relevant extensions include sleep staging, event-related potentials (ERP), steady-state visually evoked potentials (SSVEP), and emotion recognition. These paradigms cover a broader spectrum of cognitive and neurophysiological processes and are widely used in both experimental neuroscience and applied BCI research. In addition, incorporating a broader set of models – including other deep learning architectures not covered in this work – would provide a stronger foundation for comparison and allow to assess how model design choices interact with different EEG decoding tasks and data regimes.

Bibliography

[1] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "Bendr: using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," 2021. [Online]. Available: https://arxiv.org/abs/2101.12037

[2] W. Cui, W. Jeong, P. Thölke, T. Medani, K. Jerbi, A. A. Joshi, and R. M. Leahy, "Neuro-gpt: Towards a foundation model for eeg," 2024. [Online]. Available: https://arxiv.org/abs/2311.03764

[3] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu, "Large brain model for learning generic representations with tremendous eeg data in bci," 2024. [Online]. Available: https://arxiv.org/abs/2405.18765

[4] B. Aristimunha, I. Carrara, P. Guetschel, S. Sedlar, P. Rodrigues, J. Sosulski, D. Narayanan, E. Bjareholt, B. Quentin, R. T. Schirrmeister, 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," 2023. [Online]. Available: https://github.com/NeuroTechX/moabb

[5] M. Teplan, "Fundamental of eeg measurement," *MEASUREMENT SCIENCE REVIEW*, vol. 2, 01 2002.

[6] H. H. Jasper, "Report of the committee on methods of clinical examination in electroencephalography: 1957," *Electroencephalography and Clinical Neurophysiology*, vol. 10, no. 2, pp. 370–375, 1958. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0013469458900531

[7] S. J. Smith, "Eeg in the diagnosis, classification, and management of patients with epilepsy," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 76, no. Suppl 2, pp. ii2–ii7, Jun. 2005.

[8] I. Kiral-Kornek, S. Roy, E. Nurse, B. Mashford, P. Karoly, T. Carroll, D. Payne, S. Saha, S. Baldassano, T. O'Brien, D. Grayden, M. Cook, D. Freestone, and S. Harrer, "Epileptic seizure prediction using big data and deep learning: Toward a mobile system," *EBioMedicine*, vol. 27, pp. 103–111, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235239641730470X

[9] J. N. Ianof and R. Anghinah, "Traumatic brain injury: An eeg point of view," *Dementia & Neuropsychologia*, vol. 11, no. 1, pp. 3–5, Jan–Mar 2017.

[10] N. Jackson, S. R. Cole, B. Voytek, and N. C. Swann, "Characteristics of waveform shape in parkinson's disease detected with scalp electroencephalography," *eNeuro*, vol. 6, no. 3, pp. ENEURO.0151–19.2019, Jun. 2019.

[11] F. L. Teixeira, M. R. E. Costa, J. P. Abreu, M. Cabral, S. P. Soares, and J. P. Teixeira, "A narrative review of speech and eeg features for schizophrenia detection: Progress and challenges," *Bioengineering (Basel)*, vol. 10, no. 4, p. 493, Apr. 2023.

[12] K. Pan, L. Li, L. Zhang, S. Li, Z. Yang, and Y. Guo, "A noninvasive bci system for 2d cursor control using a spectral-temporal long short-term memory network," *Frontiers in Computational Neuroscience*, vol. 16, 2022. [Online]. Available: https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2022.799019

[13] M. F. Ansari, D. R. Edla, S. Dodia, and V. Kuppili, "Brain-computer interface for wheelchair control operations: An approach based on fast fourier transform and on-line sequential extreme learning machine," *Clinical Epidemiology and Global Health*, vol. 7, no. 3, pp. 274–278, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213398418301763

[14] S. Saha and M. Baumert, "Intra- and inter-subject variability in eeg-based sensorimotor brain computer interface: A review," *Frontiers in Computational Neuroscience*, vol. 13, 2020. [Online]. Available: https://www.frontiersin.org/journals/computational-neuroscience/articles/10.3389/fncom.2019.00087

[15] A. Apicella, P. Arpaia, G. D'Errico, D. Marocco, G. Mastrati, N. Moccaldi, and R. Prevete, "Toward cross-subject and cross-session generalization in eeg-based emotion recognition: Systematic review, taxonomy, and methods," *Neurocomputing*, vol. 604, p. 128354, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224011251

[16] M. Hassani and M. R. Karami, "Noise estimation in electroencephalogram signal by using volterra series coefficients," *Journal of Medical Signals and Sensors*, vol. 5, no. 3, pp. 192–200, Jul-Sep 2015.

[17] N. Faadhilah Afif, S. Harke Pratama, F. Haryanto, S. Nurul Khotimah, and Suprijadi, "Comparison of wet and dry eeg electrodes based on brain signals characterization in temporal and anterior frontal areas using audio stimulation," *Journal of Physics: Conference Series*, vol. 1505, no. 1, p. 012069, mar 2020. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/1505/1/012069

[18] Z. Koles, "The quantitative extraction and topographic mapping of the abnormal components in the clinical eeg," *Electroencephalography*

and Clinical Neurophysiology, vol. 79, no. 6, pp. 440–447, 1991. [Online]. Available: https://www.sciencedirect.com/science/article/pii/001346949190163X

[19] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (fbcsp) in brain-computer interface," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 2390–2397.

[20] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a riemannian-based kernel for bci applications," *Neurocomputing*, vol. 112, pp. 172–178, 2013, advances in artificial neural networks, machine learning, and computational intelligence. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231213001574

[21] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, apr 2018. [Online]. Available: https://dx.doi.org/10.1088/1741-2552/aab2f2

[22] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Jul. 2018. [Online]. Available: http://dx.doi.org/10.1088/1741-2552/aace8c

[23] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23730

[24] X. Wei, A. A. Faisal, M. Grosse-Wentrup, A. Gramfort, S. Chevallier, V. Jayaram, C. Jeunet, S. Bakas, S. Ludwig, K. Barmpas, M. Bahri, Y. Panagakis, N. Laskaris, D. A. Adamos, S. Zafeiriou, W. C. Duong, S. M. Gordon, V. J. Lawhern, M. Śliwowski, V. Rouanne, and P. Tempczyk, "2021 beetl competition: Advancing transfer learning for subject independence heterogenous eeg data sets," 2022. [Online]. Available: https://arxiv.org/abs/2202.12950

[25] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "4-class motor imagery," https://www.bbci.de/competition/iv/, 2008, bCI Competition IV: Dataset 2a, Graz University of Technology.

[26] R. Leeb, C. Brunner, G. Müller-Putz, A. Schlögl, and G. Pfurtscheller, "motor imagery," https://www.bbci.de/competition/iv/, 2008, bCI Competition IV: Dataset 2b, Graz University of Technology.

[27] W. Yi, S. Qiu, K. Wang, H. Qi, L. Zhang, P. Zhou, F. He, and D. Ming, "Evaluation of eeg oscillatory patterns and cognitive process during simple and compound limb motor imagery," *PLOS ONE*, vol. 9, no. 12, pp. 1–19, 12 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0114853

[28] H. Cho, M. Ahn, S. Ahn, M. Kwon, and S. Jun, "Eeg datasets for motor imagery brain-computer interface," *Gigascience*, vol. 6, no. 7, pp. 1–8, 2017. [Online]. Available: https://doi.org/10.1093/gigascience/gix034

[29] H. Liu and X. Lv, "EEG datasets of stroke patients," 12 2022. [Online]. Available: https://figshare.com/articles/dataset/EEG_datasets_of_stroke_patients/21679035

[30] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, p. 5391–5420, Aug. 2017. [Online]. Available: http://dx.doi.org/10.1002/hbm.23730

[31] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw, "Bci2000: a general-purpose brain-computer interface (bci) system," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 1034–1043, 2004.

[32] B. Zhou, X. Wu, Z. Lv, L. Zhang, and X. Guo, "A fully automated trial selection method for optimization of motor imagery based brain-computer interface," *PLOS ONE*, vol. 11, no. 9, pp. 1–20, 09 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0162657

[33] M. Kaya, M. K. Binli, E. Ozbay, H. Yanar, and Y. Mishchenko, "A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces," *Scientific Data*, vol. 5, p. 180211, Oct. 2018.

[34] A. Barachant, "Commande robuste d'un effecteur par une interface cerveau machine eeg asynchrone," Ph.D. dissertation, Université de Grenoble, 2012. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01196752

[35] J. Faller, C. Vidaurre, T. Solis-Escalante, C. Neuper, and R. Scherer, "Autocalibration and recurrent adaptation: Towards a plug and play online erd-bci," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 3, pp. 313–319, 2012.

[36] R. Scherer, J. Faller, E. V. C. Friedrich, E. Opisso, U. Costa, A. Kübler, and G. R. Müller-Putz, "Individually adapted imagery improves

brain-computer interface performance in end-users with disability," *PLOS ONE*, vol. 10, no. 5, pp. 1–14, 05 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0123727

[37] J. F. Cavanagh, A. Napolitano, C. Wu, and A. Mueen, "The patient repository for eeg data + computational tools (pred+ct)," *Frontiers in Neuroinformatics*, vol. 11, 2017. [Online]. Available: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2017.00067

[38] I. Obeid and J. Picone, "The temple university hospital eeg data corpus," *Frontiers in Neuroscience*, vol. 10, 2016. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00196

[39] J. F. Cavanagh, P. Kumar, A. A. Mueller, S. P. Richardson, and A. Mueen, "Diminished eeg habituation to novel events effectively classifies parkinson's patients," *Clinical Neurophysiology*, vol. 129, no. 2, pp. 409–418, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245717311719

[40] M. A. Albrecht, J. A. Waltz, J. F. Cavanagh, M. J. Frank, and J. M. Gold, "Increased conflict-induced slowing, but no differences in conflict-induced positive or negative prediction error learning in patients with schizophrenia," *Neuropsychologia*, vol. 123, pp. 131–140, 2019, cognitive Effort. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0028393218301726

[41] 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*, vol. 117, pp. 113–122, Aug. 2018, epub 2018 May 23.

[42] D. R. Brown, S. P. Richardson, and J. F. Cavanagh, "An eeg marker of reward processing is diminished in parkinson's disease," *Brain Research*, vol. 1727, p. 146541, Jan. 2020, epub 2019 Nov 5.

[43] 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*, vol. 47, no. 8-9, pp. 1978–1987, Jul. 2009, epub 2009 Mar 17.

[44] 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*, vol. 132, p. 107125, Sep. 2019, epub 2019 Jun 19.

[45] A. Singh, R. C. Cole, A. I. Espinoza, D. Brown, J. F. Cavanagh, and N. S. Narayanan, "Frontal theta and beta oscillations during lower-limb movement in parkinson's disease," *Clinical Neurophysiology*,

vol. 131, no. 3, pp. 694–702, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245720300092

[46] A. Singh, R. C. Cole, A. I. Espinoza, A. Evans, S. Cao, J. F. Cavanagh, and N. S. Narayanan, "Timing variability and midfrontal ∼4hz rhythms correlate with cognition in parkinson's disease," *NPJ Parkinson's Disease*, vol. 7, no. 1, p. 14, Feb. 2021, published 2021 Feb 15.

[47] L. Veloso, J. McHugh, E. von Weltin, S. Lopez, I. Obeid, and J. Picone, "Big data resources for eegs: Enabling deep learning research," 12 2017, pp. 1–3.

[48] A.-K. Kiessner, R. T. Schirrmeister, L. A. Gemein, J. Boedecker, and T. Ball, "An extended clinical eeg dataset with 15,300 automatically labelled recordings for pathology decoding," *NeuroImage: Clinical*, vol. 39, p. 103482, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213158223001730

[49] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[50] G. King and L. Zeng, "Logistic regression in rare events data," *Political Analysis*, vol. 9, 09 2002.