Towards a foundation model for EEG data

Angela Hu
McGill University
COMP396 Research Report
qingchen.hu@mail.mcgill.ca

Abstract

This project, in collaboration with researchers at Mila, explores the development and benchmarking of pre-trained transformer models for decoding electroencephalography (EEG) signals, with the goal of advancing neural population dynamics analysis. While MOABB provides a robust foundation for evaluation, its methods were adapted to enable comparisons of POYO's performance by introducing the Entire-Dataset Evaluation Scheme, which offers a more comprehensive and effective evaluation tailored to POYO's assessment context.

The later part of the study evaluates LaBraM, a transformer-based EEG foundation model, on the MOABB dataset to establish baselines for comparison with models like POYO. Using the AlexMI and PhysionetMI datasets for binary motor imagery tasks, initial fine-tuning via linear probing demonstrated poor transferability, with validation metrics near random. Fine-tuning the entire model revealed overfitting due to the mismatch between MOABB's smaller, noisier datasets and the curated datasets used for pretraining. Data analysis highlighted substantial variability across EEG channels, underscoring the challenge of balancing preprocessing consistency with preserving data complexity to assess model robustness. These findings point to the need for refined fine-tuning strategies to adapt large foundation models for realistic EEG scenarios. The corresponding GitHub repositories are provided as follows: for the MOABB benchmark repository, see `https://github.com/Angelawork/EEG-Foundation-model_LiNC-Lab_COMP396`, and for the fine-tuning of LaBraM, refer to `https://github.com/RoyHEyono/LincLab_LaBraM/tree/Angela_finetune`.

1 Background and Related Work

1.1 The MOABB Benchmark

The MOABB (Mother of All BCI Benchmarks) framework addresses critical challenges in reproducibility within Brain-Computer Interface (BCI) research. By benchmarking 30 machine learning pipelines across 36 publicly available EEG datasets, MOABB emphasizes paradigms such as motor imagery [Jayaram and Barachant, 2018]. MOABB's open framework offers standardized benchmarking through uniform data retrieval, preprocessing, and cross-validation. These practices allow researchers to evaluate model performance consistently using data's subject-trial pairs, ensuring clear and unbiased comparisons of model capabilities [Aristimunha et al., 2023].

1.2 POYO Framework

POYO is an advanced framework designed for large-scale neural decoding, capable of modeling diverse neural data across various modalities. While originally developed for spiking data, POYO has also been shown to perform effectively on other modalities, such as ECoG, calcium imaging, and EEG. Leveraging a transformer-based architecture and an innovative spike tokenization strategy, POYO excels at capturing neural population dynamics across different sessions and individuals. By addressing the challenges posed by neural variability, POYO demonstrates superior capabilities in transfer learning and few-shot learning. This framework significantly outperforms traditional decoding methods, enhancing the scalability and generalizability of neural data analysis [Azabou et al., 2023].

2 Benchmark on Baseline Classifiers

Previous work on MOABB established a reproducible benchmark for evaluating EEG classification pipelines using standardized datasets. The benchmark allows clear comparison of model performance across various paradigms. Establishing benchmark scores using MOABB on model pipelines is essential for validating POYO's performance against widely accepted standards in EEG-based BCI research, particularly for motor imagery tasks.

While MOABB provides a strong foundation, its evaluation methods does not fully align with POYO's methodologies. The evaluation mechanisms in MOABB, designed for traditional pipelines, may lack the specificity needed to assess foundation models like POYO comprehensively. To address this gap, this part of the project focuses on developing a complementary evaluation scheme while re-evaluating the baseline classifiers to establish a reference for POYO's performance on MOABB datasets. This approach ensures that POYO's unique contributions are effectively contextualized, thoroughly assessed, and accurately represented within the framework of standardized benchmarking.

2.1 Datasets and Data Preparation

The primary datasets used in this project are **BNCI2014-001**, **BNCI2014-004**, and **Zhou2016**, selected due to their superior data quality compared to other available options. A flexible range of functions is provided in the `dataset_setup.py` script to enable customization for filtering and modifying the data. For example, specifying selections such as `BNCI2014-001_1_0.h5`, `BNCI2014-001_1_1.h5`, and `Zhou2016_3_2.h5` in a file allows one to filter and use specific subject-session pairs within the datasets—e.g., Subject 1's Sessions 0 and 1 from **BNCI2014-001**, and Subject 3's Session 2 from **Zhou2016**. The setup accounts for variations in naming conventions for session identifiers, providing robust functionality for precise data selection.

It is important to note that while filtering subject and session IDs can also be achieved through MOABB's compound dataset object, this approach resamples the data to a frequency of 250 Hz by default, which has been shown to degrade the performance of baseline classifiers. Consequently, the manual filtering method addressed above is recommended, particularly for identifying and excluding data from low-quality channels. Additionally, note that all data retrieved from MOABB automatically apply a 50 Hz notch filter to mitigate power-line interference.

2.2 Evaluation Schemes

MOABB provides three evaluation schemes: **Within-Session**, **Cross-Session**, and **Cross-Subject**. These evaluation schemes are based on the unique data structure implemented in MOABB's datasets, which consist of subjects, each having multiple sessions, with each session containing multiple trials corresponding to distinct neural recordings. The data is preprocessed to retain only the relevant portions, from the start of the trial to the rest period, ensuring alignment with the desired labels for downstream tasks. The three different schemes enable evaluation at different levels, each assessing the model's ability to generalize across varying levels of session and subject variability.

Within-Session Evaluation This scheme trains and tests a model on data from the same session, splitting the data into training and testing subsets. Trials are shuffled before k-fold cross-validation to assess generalization performance while minimizing overfitting within a single session.

Cross-Session and Cross-Subject Evaluations These evaluations employ a leave-one-out cross-validation approach, designating one session or subject, respectively, as the test dataset while training the model on the remaining data. These methods emphasize transfer learning by addressing variability across subjects and sessions.

Entire-Dataset Evaluation While the standard schemes are effective for traditional pipelines, they differ from POYO's approach in several key aspects. To align the evaluation process with POYO's evaluation methodology, this project introduces the **Entire-Dataset Evaluation Scheme**. This approach leverages data from all subjects and sessions within a dataset.

Entire-Dataset Evaluation Workflow:

- **Outer 5-Fold Cross-Validation:**
  - *Training Data:* 80% of the shuffled data from all subjects and sessions.

  - *Test Data:* The remaining 20% of the data.
- **Stratified Inner 3-Fold Cross-Validation (on the Outer Training Data):**
  - *Inner Training Data:* Two-thirds of the outer training data.
  - *Inner Validation Data:* One-third of the outer training data. This step fine-tunes hyperparameters by evaluating different configurations.
- **Model Training and Evaluation:**
  - After hyperparameter tuning, the model is trained on the entire outer training set.
  - Testing is conducted on the held-out 20% test set, yielding five scores (one per fold) rather than scores tied to individual subject-session pairs.

This comprehensive evaluation scheme was integrated into MOABB's benchmarking function, allowing model pipelines to be specified through YAML configuration files.

2.3 Experiment setup

To ensure uniformity and consistency in evaluating different models, the pipeline.py script offers a flexible approach, enabling seamless switching between cognitive task paradigms, datasets, and evaluation schemes. This flexibility is controlled by a random seed for reproducibility. For this project, the experiments focused exclusively on the Left vs. Right Hand motor imagery paradigm. All pipelines available in MOABB's benchmark table, such as ACM+TS+SVM, CSP+SVM, and EEGITNet, were utilized for assessment. These pipelines were either strictly initialized as per the MOABB repository's specifications or configured using YAML files provided by the framework. The default evaluation metric employed was AUROC: area under the receiver operating characteristic curve, suitable for binary classification tasks.

The evaluation's results are averaged across all cross-validation folds, and the final benchmark results were reported as mean scores and standard deviations over 5 seeds to reflect the model's performance consistency and reliability.

3 Benchmark Results

The benchmarking revealed that most evaluation results closely align with MOABB's within-session evaluation scores, demonstrating consistency with the baseline setup.

Table 1: Entire Dataset Benchmark Results

pipeline	BNCI2014-001	BNCI2014-004	Zhou2016
CSP+LDA	77.61±2.47	76.26±1.05	91.59±1.21
CSP+SVM	81.52±1.97	77.84±0.87	93.0±1.28
DLCSPauto+shLDA	77.61±2.48	76.26±1.05	91.64±1.22
FgMDM	83.08±1.37	76.16±1.06	93.52±1.56
LogVariance+LDA	73.78±1.78	75.36±1.02	90.05±1.36
LogVariance+SVM	73.54±1.77	75.43±1.0	90.22±1.44
MDM	62.13±4.24	75.57±1.13	86.8±4.06
TRCSP+LDA	75.25±3.0	76.28±1.04	91.78±1.62
TS+EL	84.75±1.29	76.24±1.04	94.24±1.3
TS+LR	84.72±1.31	76.25±1.04	94.31±1.3
TS+SVM	88.71±0.8	79.39±0.9	94.77±1.01
Keras_EEGITNet	88.45±1.3	74.88±1.84	97.19±0.81
Keras_EEGNeX	83.59±1.42	71.85±1.36	95.56±0.94
Keras_EEGNet_8_2	89.26±0.94	76.99±1.01	96.68±0.63
Keras_EEGTCNet	90.0±0.35	76.75±0.84	96.36±0.78
Keras_ShallowConvNet	91.97±1.34	75.32±0.52	96.58±1.00

4 LaBraM Fine-Tuning

The MOABB framework does not provide baseline results for any EEG foundation models, which makes it necessary to establish such baselines for validation purposes. In this context, the goal was to define a baseline for

Table 2: Cross Subject Benchmark Results

pipeline	BNCI2014-001	BNCI2014-004	Zhou2016
CSP+LDA	76.16±16.05	79.37±14.29	92.22±5.64
CSP+SVM	75.95±15.92	73.72±17.14	91.07±5.99
DLCSPauto+shLDA	76.17±16.07	79.37±14.29	92.37±5.78
FgMDM	76.14±15.57	78.43±14.36	91.23±5.93
LogVariance+LDA	69.77±15.27	78.56±14.24	89.03±6.60
LogVariance+SVM	69.60±15.24	78.58±14.26	89.15±6.82
MDM	75.95±12.47	78.92±14.51	93.62±5.17
TRCSP+LDA	73.73±14.23	78.67±14.32	92.35±5.67
TS+EL	77.73±15.72	78.57±14.33	92.41±5.78
TS+LR	77.65±15.70	78.58±14.32	92.32±5.55
TS+SVM	76.35±16.43	71.92±16.64	91.50±5.56
Keras_EEGITNet	85.20±8.73	75.72±12.56	94.66±2.84
Keras_EEGNeX	76.42±6.57	65.70±12.80	92.41±4.65
Keras_EEGNet_8_2	82.45±13.19	71.30±15.61	94.62±3.55
Keras_EEGTCNet	82.82±12.84	70.53±15.67	94.74±2.61

Table 3: Within Session Benchmark Results

pipeline	BNCI2014-001	BNCI2014-004	Zhou2016
CSP+LDA	82.66±16.68	80.11±15.28	93.4±6.98
CSP+SVM	82.6±16.7	79.14±15.99	93.47±7.18
DLCSPauto+shLDA	82.74±16.58	79.95±15.32	92.88±7.15
FgMDM	86.33±13.0	79.39±15.48	92.64±6.63
LogVariance+LDA	77.69±16.11	78.86±15.01	88.54±10.13
LogVariance+SVM	76.34±17.17	78.48±15.38	88.32±8.43
MDM	80.83±16.43	77.72±16.1	90.43±7.22
TRCSP+LDA	79.66±16.71	79.6±15.71	93.17±7.25
TS+EL	86.33±13.93	80.0±15.38	94.72±5.91
TS+LR	86.96±13.39	80.18±15.29	94.56±5.87
TS+SVM	86.47±14.1	79.45±15.75	93.59±6.27
Keras_EEGITNet	76.25±15.46	66.80±15.98	67.91±14.64
Keras_EEGNeX	69.98±16.11	66.40±17.51	63.62±16.59
Keras_EEGNet_8_2	74.79±21.01	69.52±19.21	92.35±8.78
Keras_EEGTCNet	60.76±17.40	62.00±18.54	77.05±11.85

foundation models to enable the evaluation of progress made with POYO. To achieve this, it was decided to first train LaBraM in a supervised learning setting and then apply the Entire-Dataset Evaluation Scheme to ensure a thorough evaluation.

Base Model LaBraM is based on the Transformer architecture, where raw EEG signals are first segmented into fixed-length patches. Each patch is processed with a temporal encoder to extract features, which are then enriched with temporal and spatial embeddings. The sequence of embeddings is passed through a Transformer encoder using patch-wise attention to generate the final output, capturing both temporal and spatial information of the EEG data [Jiang et al., 2024].

Pre-training The model was pre-trained using a masked modelling approach, where random patches of EEG signals were masked, and the model predicted the missing tokens. This pretraining utilized over 2,500 hours of EEG data across diverse datasets, allowing the model to generalize effectively across different EEG configurations and tasks.

5 Experiment

The goal of this experiment is to pretrain, fine-tune, and evaluate LaBraM on the MOABB dataset, establishing it as a reference foundation model for POYO. The primary datasets of focus are AlexMI and PhysionetMI, specifically for the binary classification task of Right Hand vs. Feet motor imagery. Upon inspection, the 16-channel AlexMI and 64-channel PhysionetMI datasets are fully compatible with the standard 10-20 EEG electrode system, which is the channel configuration accepted by LaBraM.

- **AlexMI Channel Names:** FPZ, F7, F3, FZ, F4, F8, T7, C3, CZ, C4, T8, P7, P3, PZ, P4, P8.
- **PhysionetMI Channel Names:** FC5, FC3, FC1, FCZ, FC2, FC4, FC6, C5, C3, C1, CZ, C2, C4, C6, CP5, CP3, CP1, CPZ, CP2, CP4, CP6, FP1, FPZ, FP2, AF7, AF3, AFZ, AF4, AF8, F7, F5, F3, F1, FZ, F2, F4, F6, F8, FT7, FT8, T7, T8, T9, T10, TP7, TP8, P7, P5, P3, P1, PZ, P2, P4, P6, P8, PO7, PO3, POZ, PO4, PO8, O1, OZ, O2, IZ.

5.1 Pre-training Stage

The raw data for **AlexMI** and **PhysionetMI** are initially provided in the `.fif` and `.edf` formats, respectively. To ensure consistent processing across datasets, the data are pre-processed as follows:

- A random seed is used to isolate 50% of the subjects for both the pre-training and fine-tuning stages.
- Only channels conforming to the standard 10-20 system, along with a few additional channels specified by LaBraM, are included.
- The data is band-pass filtered between 8-32 Hz to capture the relevant frequency range for motor imagery and resampled to a rate of 200 Hz. These preprocessing steps are essential for preparing the data for model training, ensuring consistency across datasets.

5.2 Fine-tuning Stage

For the fine-tuning stage, the data preprocessing steps were kept minimal to align closely with POYO and LaBraM's preprocessing pipeline. The EEG signals were first retrieved through the MOABB interface, specifying the downstream task as binary classification for right-hand versus feet motor imagery. Note that all MOABB data have been pre-processed with a 50 Hz notch filter already.

- The raw EEG signals were then converted into `MNE Epochs` objects and filtered between 8 Hz and 32 Hz to remove low-frequency noise.
- The data were subsequently resampled to 200 Hz to match LaBraM's specifications.
- Outlier removal for each single data point per session and per channel was conducted by clipping the data at the 1st and 99th percentiles. Subsequently, normalization across subjects was performed using **min-max scaling**, which scaled the data to fall within the range of -0.1 to 0.1.

Finally, the data were split into training, validation, and test sets using a 60% training, 30% validation, and 10% test ratio. The split was performed equally across subject-session pairs, ensuring that the model was exposed to data from multiple subjects, thereby improving its generalization across different EEG configurations.

5.3 Dataset Status

Multiple approaches were explored to examine and compare data trends after preprocessing, given the significant differences between the datasets used to train the original LaBraM model (e.g., TUAB and TUEV) and the MOABB dataset targeted for assessment. After concatenating data for each subject across trials and sessions, the minimum, maximum, and average values were computed across EEG channels and visualized on separate axes. The figures 1 and 2 below, based on randomly selected subjects (from a total of 109 subjects in PhysionetMI), demonstrate substantial variability in the data characteristics after preprocessing.

The minimum and maximum values exhibit large fluctuations across EEG channels, while the average values reveal a more consistent trend. However, some channels show notable deviations from this consistency. These patterns underscore the need for careful preprocessing to ensure data uniformity and comparability across subjects and sessions. Additionally, they emphasize the importance of evaluating the contribution of individual EEG channels to the overall signal.

Despite this, minimizing preprocessing is equally critical for assessing the model's robustness to noisy and varied data, as MOABB represents a more realistic and "in-the-wild" dataset compared to the highly curated TUAB and TUEV datasets. The variability in MOABB data reflects conditions closer to real-world EEG applications, making it particularly valuable for testing model adaptability and generalization under practical circumstances. This highlights a trade-off: ensuring preprocessing consistency versus preserving data complexity to evaluate the model's ability to handle realistic noise and variability effectively.

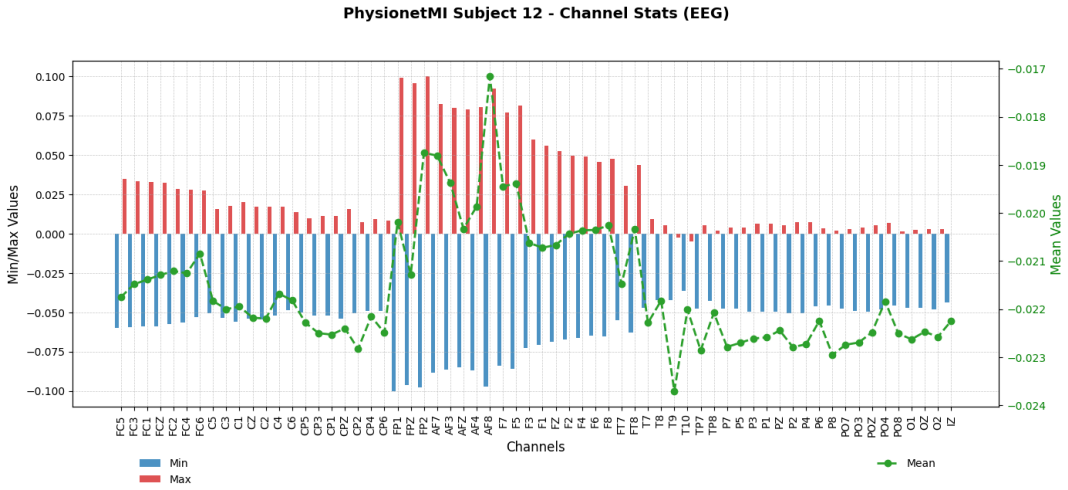

Figure 1: Channel-wise Min,Max and Mean for preprocessed subject 12 in PhysionetMI

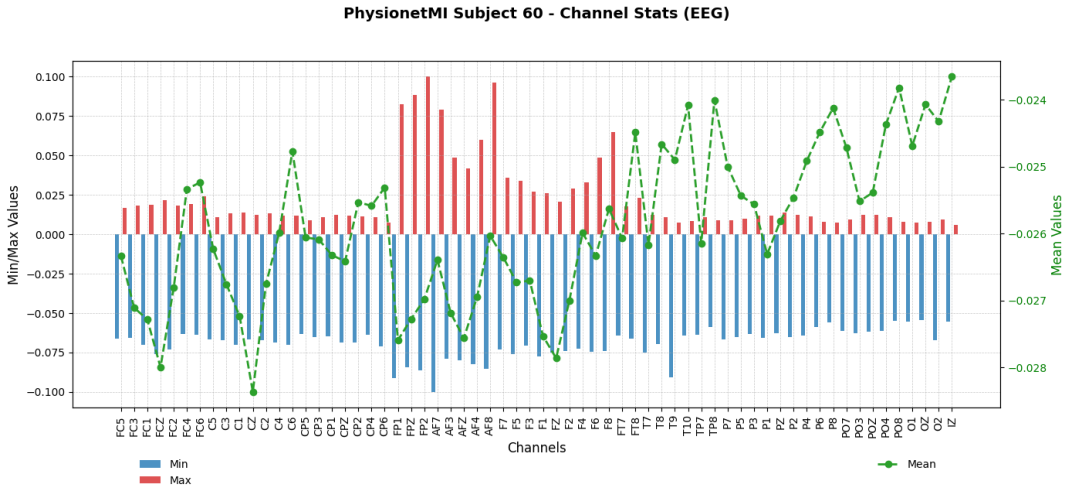

Figure 2: Channel-wise Min,Max and Mean for preprocessed subject 60 in PhysionetMI

6 LaBraM Fine-Tuning Results and Discussion

Fine-Tuning with Linear Probing After pretraining the model on both the AlexMI and PhysionetMI datasets, we attempted fine-tuning by freezing the pretrained model parameters and training only the linear classification layer for the downstream task. However, as shown in figure 3, this approach failed to achieve satisfactory performance. The training metrics showed improvement, with decreasing training loss and increasing training accuracy, yet validation metrics remained poor. Specifically, validation loss remained high, and validation accuracy showed erratic fluctuations without meaningful improvement. Moreover, the validation AUROC stagnated at approximately 0.5, indicating performance no better than random chance. These results suggest that the features

learned during pretraining are not adequately transferable to the downstream task when the model is frozen. The lack of improvement in validation performance indicates that freezing the model limits its adaptability to the specific characteristics of the new dataset.

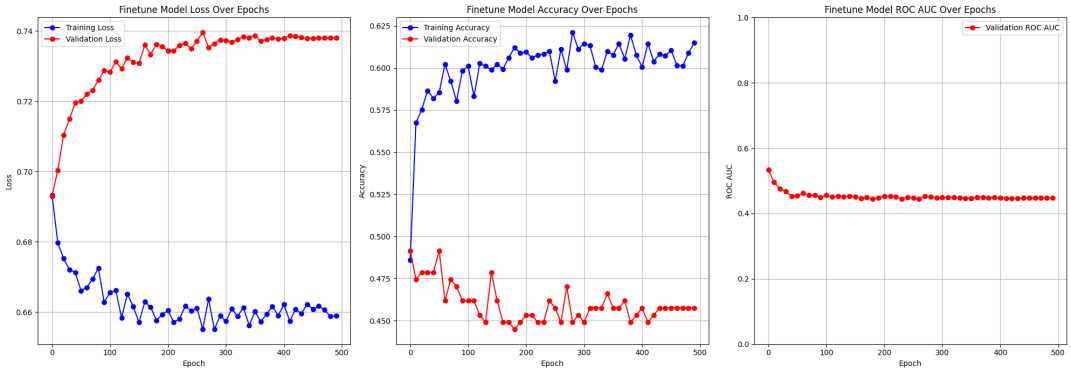

Figure 3: Fine-Tuning with Linear Probing

Fine-Tuning the Entire Model Subsequently, we attempted fine-tuning the entire model by using all parameters. Unfortunately, this approach revealed a significant overfitting issue in figure 4. The training loss rapidly converged to near zero, and training accuracy approached perfect performance, but the validation loss steadily increased. Validation accuracy remained stagnant, hovering around 0.5, and the validation AUROC fluctuated near 0.55, reflecting poor generalization. These observations indicate that the large pretrained LaBraM model, initially trained on extensive datasets like TUAB and TUEV, struggles to adapt to the relatively small MOABB dataset. The model's high capacity appears to be excessive for the limited data, resulting in overfitting to the training set while failing to generalize to unseen data. These findings underscore the need for techniques such as regularization, data augmentation, or transfer learning adaptations to better balance the model's complexity with the scale of the available dataset.

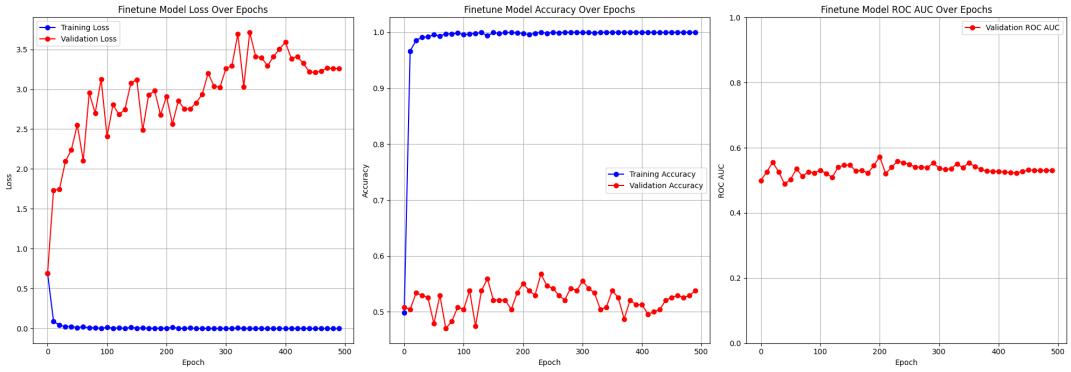

Figure 4: Fine-Tuning the Entire Model

Fine-Tuning the Entire Model: Preprocessing Adjustments and Hyperparameter-Tuning A grid search was conducted to identify the optimal set of hyperparameters, and various combinations of preprocessing techniques were tested. The results revealed that maintaining all previous preprocessing steps, but stopping after outlier removal without performing min-max scaling, yielded significantly better performance. The results highlight the model's capacity to achieve enhanced performance when appropriately tuned with carefully selected preprocessing methods and hyperparameters. Plots in figure 5 demonstrate that overfitting has been significantly reduced after the search. In the accuracy plot, the training accuracy stabilizes near 0.99, while validation and test accuracies maintain consistent levels around 0.75–0.78, with minimal gaps between them, indicating strong generalization. Similarly, in the AUROC plot, the validation and test AUROC curves remain closely aligned, stabilizing at high values (0.82–0.85).

Table 4: Hyperparameters with their possible values and optimal configurations.

Hyperparameter	Searched Values	Optimal Value
Learning rate	[1e-4, 1e-3, 1e-2]	1e-4
Weight decay	[0.5, 0.05, 1e-4]	0.5
DropOut	[0, 1e-3, 1e-4]	0
Layer-wise decay	[0.65, 0.05, 1e-4]	0.65
Batch size	[16, 32, 64]	16
Drop path	[0.01, 0.1, 1e-4]	0.01

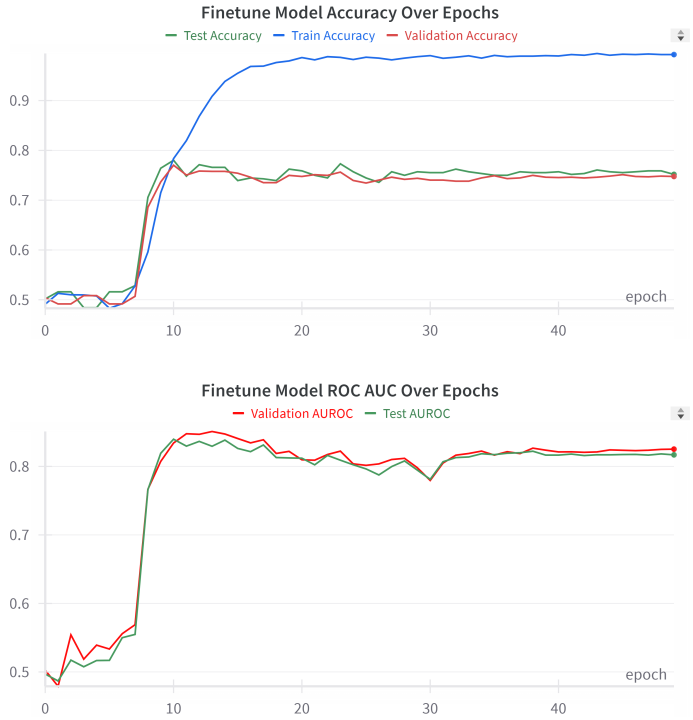

Figure 5: Model Performance with Optimal Hyperparameter Setting

References

Bruno Aristimunha, Igor Carrara, Pierre Guetschel, Sara Sedlar, Pedro Rodrigues, Jan Sosulski, Divyesh Narayanan, Erik Bjareholt, Barthelemy Quentin, Robin Tibor Schirrmeister, Emmanuel Kalunga, Ludovic Darmet, Cattan Gregoire, Ali Abdul Hussain, Ramiro Gatti, Vladislav Goncharenko, Jordy Thielen, Thomas Moreau, Yannick Roy, Vinay Jayaram, Alexandre Barachant, and Sylvain Chevallier. Mother of all BCI Benchmarks, 2023. URL `https://github.com/NeuroTechX/moabb`.

Mehdi Azabou, Vinam Arora, Venkataramana Ganesh, Ximeng Mao, Santosh Nachimuthu, Michael Mendelson, Blake Richards, Matthew Perich, Guillaume Lajoie, and Eva L. Dyer. A unified, scalable framework for neural population decoding. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Vinay Jayaram and Alexandre Barachant. Moabb: trustworthy algorithm benchmarking for bcis. *Journal of neural engineering*, 15(6):066011, 2018.

Weibang Jiang, Liming Zhao, and Bao liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=QzTpTRVtrP`.