ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institut für Integrierte Systeme
Integrated Systems Laboratory

DEPARTMENT OF
INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Fall Semester 2024

Deep Learning Based EEG Signal Understanding using Foundational Models

Master Thesis

Nicolas Lehmann
lehmannn@student.ethz.ch

15.04.2025

Advisors: Thorir Mar Ingolfsson, thoriri@iis.ee.ethz.ch
Dr. Yawei Li, yawei.li@vision.ee.ethz.ch

Professor: Prof. Dr. Luca Benini, lbenini@iis.ee.ethz.ch

Abstract

Electroencephalography (EEG) provides a non-invasive window into brain activity with critical clinical and research applications such as seizure detection, sleep-stage analysis, and brain–computer interfaces. However, EEG signals are complex—they exhibit low signal-to-noise ratios, limited spatial resolution, and high inter-subject variability—which makes extracting robust and generalizable features a challenging task. Motivated by the need to leverage large volumes of unlabelled EEG data and reduce the reliance on costly annotations, this project investigates self-supervised learning (SSL) as a strategy for automatic EEG representation learning.

The objectives are threefold: (1) to explore diverse SSL pretraining methods, including masked reconstruction (with techniques such as random and clustered patch masking as well as lowpass-filtered target reconstruction) and contrastive learning (using frequency-domain, topological channel-mixing, and masking-based augmentations); (2) to evaluate these strategies across multiple model architectures—namely MEST (a transformer-based encoder with alternating attention), FEMBA (a scalable state-space model), and LaBraM (a well-established EEG foundation model with public weights); and (3) to assess the impact of these approaches on downstream EEG classification tasks using three diverse clinical datasets.

Key challenges include managing outliers, severe class imbalances, and addressing the computational complexity of processing long, multi-channel recordings. This work combines carefully designed SSL objectives with varying EEG-specific model architectures, evaluated on three downstream classification tasks. Extensive experiments — both through direct finetuning and via evaluations of frozen encoder embeddings — demonstrate that optimized SSL pretraining techniques can outperform the current standard of Random Masked Reconstruction. Specifically, using *Contrastive Learning with Channel-Mixing*, the AUROC performance of FEMBA on the TUSL dataset has been improved by +4.9% compared to the original work.

Acknowledgments

I sincerely thank my supervisors Thorir Mar Ingolfsson and Dr. Yawei Li for their invaluable advice and support throughout this work. I am also grateful to Prof. Benini and the IIS lab for providing the opportunity and resources to conduct this research. Lastly, I appreciate the support and collaboration of my peers at the IIS lab, whose contributions and shared moments made this experience both productive and enjoyable.

Declaration of Originality

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. **In consultation with the supervisor**, one of the following two options must be selected:

☐ I hereby declare that I authored the work in question independently, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].

☒ I hereby declare that I authored the work in question independently. In doing so I only used the authorised aids, which included suggestions from the supervisor regarding language and content and generative artificial intelligence technologies. The use of the latter and the respective source declarations proceeded in consultation with the supervisor.

Title of paper or thesis:

Deep Learning Based EEG Signal Understanding using Foundational Models

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Lehmann	Nicolas

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guidelines.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zürich, 07.04.2025	

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] For further information please consult the ETH Zurich websites, e.g. https://ethz.ch/en/the-eth-zurich/education/ai-in-education.html and https://library.ethz.ch/en/researching-and-publishing/scientific-writing-at-eth-zurich.html (subject to change).

Contents

Chapter 1

Introduction

Electroencephalography (EEG) offers a unique, non-invasive window into the human brain through the measurement of electrical activity. Its clinical applications span from seizure detection to sleep-stage identification [1, 2], making it valuable in both diagnostic and research settings. However, despite its transformative potential, EEG remains a challenging signal to work with due to its inherent low signal-to-noise ratio, limited spatial resolution and high inter-subject variability.

In recent years, advancements in deep learning, both architectural and methodical, have transformed domains such as language and image processing [3, 4, 5], opening new avenues for their application on complex signals such as EEG. The success of large-scale transformer models in language and vision has demonstrated that models pretrained on vast unlabelled datasets can later be finetuned effectively for downstream tasks with superior performance compared to pure supervised training. EEG, with a scarcity of annotated data but large availability of unlabelled recordings, stands to benefit considerably from such self-supervised learning (SSL) approaches.

Existing research in EEG foundation models has largely focused on architectural adaptations. Researchers have explored various transformer-based designs, with custom attention functions to adapt to EEG-specific spatio-temporal characteristics [6], or state-space models tailored to handle long-sequence data in an efficient manner [7]. Most work to date employs either masked reconstruction or, less frequently, contrastive methods. Yet, a systematic exploration of dedicated SSL training strategies specific to EEG signal data remains an open research question.

This work addresses that gap by exploring, developing and evaluating a range of SSL pretraining strategies specifically designed for EEG. Both masked reconstruction and contrastive learning methods are systematically investigated across three distinct model architectures. These models are pretrained on a large corpus of unlabelled EEG data and subsequently evaluated by finetuning on three different EEG classification tasks. In doing so, this work aims to answer the following research questions:

1. **How can SSL techniques be leveraged for meaningful representation learning in EEG?** Which pretraining strategy leads to superior downstream performances?

How do they affect the quality of embeddings? What challenges arise in the transfer of SSL methodologies from vision and language to EEG?

2. **How well do SSL strategies apply to different model architectures and downstream tasks?** Are there differences in which methods are most effective between the model architectures, and which methods promise to be the most consistent across varied tasks?

3. **What specific challenges do the pretraining and finetuning datasets present, and how can these challenges be effectively addressed?** In particular, mitigating issues related to large outliers, imbalanced class distributions and overfitting are crucial for developing effective EEG analysis models.

The main contributions of this thesis are as follows:

- **Novel Pretraining Strategies:** A variety of SSL strategies are explored that extend beyond standard masked reconstruction and contrastive approaches. In a comparative study, different masking schemes as well as EEG specific signal preprocessing and augmentations for contrastive learning are proposed.

- **Comprehensive Evaluation:** The proposed SSL methods are evaluated from two perspectives. First, by finetuning the pretrained models on diverse EEG downstream tasks, and second, through both qualitative and quantitative analysis of the embedding quality of the learned representations without finetuning.

- **Insights into Model and Downstream Tasks Adaptation:** By comparing three different model architectures, understanding of the influence of architectural choice on the effectiveness of the SSL strategy is gained. Differences of applicability of the methods to the three downstream tasks are revealed.

The remainder of this report is structured as follows. In Chapter 2 (Background), we review the fundamentals of EEG, discuss the clinical and research significance of the signal, and provide an overview of foundation models and SSL techniques. Chapter 3 (Related Work) surveys the current landscape of EEG foundation models and SSL strategies, pinpointing the research gap that this work addresses. Chapter 4 (Implementation) details the various model architectures and pretraining strategies developed for this study, as well as an overview of additional explorations and various functional implementations necessary for the model trainings . Chapter 5 (Results) presents both quantitative and qualitative evaluations of the methods across multiple EEG downstream tasks, along with an analysis of the challenges encountered during pretraining and fine-tuning. Finally, Chapter 6 (Conclusion and Future Work) summarizes the key findings and outlines directions for further research.

In summary, this thesis investigates the potential of SSL techniques to unlock more robust and transferable representations from EEG data. By leveraging large unlabelled datasets and exploring a diverse set of pretraining strategies, this work aims to offer new insights into how these methods can be effectively applied to deep-learning for EEG signals in clinical and research settings.

Chapter 2

Background

2.1. Electroencephalography

Electroencephalography (EEG) is a non-invasive technique to record the electrical activity of the brain. EEG signals are typically recorded using electrodes placed on the scalp of the patient according to standardized positioning systems, with the International 10-20 system [8] being the most widely adopted. Recordings may be configured as either unipolar or bipolar montages. In unipolar recordings, each electrode's potential is measured against a common reference electrode or the reference average. Bipolar recordings measure the potential differences between pairs of electrodes, which can emphasize local activity while reducing common artifacts. These representations are generally mathematically convertible; this work primarily uses bipolar measurements, performing the necessary conversion where unipolar datasets demand it.

EEG signals typically contain frequency components primarily in the range between 1 and 100 Hz. These frequencies are conventionally divided into distinct bands that are associated with different cognitive states: delta (<4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (>30 Hz). [9]

In a clinical setting, EEG can serve as a crucial diagnostic tool for epilepsy [1], sleep stage detection [10] and anomaly detection [11]. Furthermore, EEG is not only used for clinical applications, but also used for other venues such as Brain-Computer Interfaces (BCI) [12] or communication assistance [13].

Furthermore, EEG is used in research settings for applications such as brain-computer interfaces (BCI) [12].

2.2. Temple University EEG Corpus

2.2.1. Datasets for Pretraining

The Temple University EEG Corpus (TUEG) [2] represents one of the largest publicly available EEG datasets, comprising approximately 26,000 recordings collected from over

14,000 patients. This extensive collection amounts to roughly 21,000 hours of EEG data, making it an ideal resource for large-scale pretraining.

The majority of recordings follow the standard 10-20 international system with 22 EEG channels and are sampled at 256 Hz. For this research, the continuous recordings were segmented into non-overlapping 5-second windows, creating manageable units for model training while preserving temporal dynamics. To ensure valid evaluation during downstream finetuning, all recordings from patients whose data appear in the finetuning datasets (TUAB, TUSL, and TUAR) were excluded from the pretraining corpus.

2.2.2. Datasets for Finetuning

For the purpose of evaluating the pretrained models, the following datasets were used for finetuning and assessing the classification performance on downstream tasks:

- **TUAB (Abnormal EEG Corpus)**: Binary classification of normal vs. abnormal EEG recordings. This dataset includes 2,329 subjects with balanced class distribution.
- **TUAR (Artifact Corpus)**: Detection and classification of various artifacts in EEG recordings (6 classes). This dataset contains 213 subjects with annotations for artifacts such as eye blinks and muscle movements.
- **TUSL (Slowing Corpus)**: Multi-class classification to of slowing events, seizures, complex background, and normal EEG (4 classes). The dataset consists of 38 subjects with a highly imbalanced class distribution.

2.3. Foundation Models

Foundation models represent a paradigm shift in machine learning, characterized by their scale, generalizability and adaptability across a wide range of tasks. As defined by [14], foundation models are large-scale models trained on vast amounts of data, often using self-supervised learning techniques. These models are designed to be finetuned for specific downstream tasks, leveraging their pretrained knowledge to achieve state-of-the-art performance across various applications. Unlike traditional task-specific models, foundation models serve as versatile building blocks that capture generalizable patterns and representations in the unlabelled data.

2.3.1. Transformer Models

Transformers [15] have emerged as the dominant architecture for foundation models across various domains. Originally developed for natural language processing tasks, this attention-based architecture has demonstrated remarkable versatility and scalability, becoming the backbone of models like BERT [4] and GPT [3].

The defining characteristic of transformer architectures is the self-attention mechanism (see section 2.3.1.1), which allows the model to weigh the importance of different parts

of the input sequence when processing each element. This enables capturing global dependencies across the entire input context, a significant advantage over previously dominant architectures: CNNs are inherently limited to local patterns through fixed receptive fields, while RNNs process information sequentially, struggling with long-range dependencies due to the vanishing gradient problem. In contrast, transformers establish direct connections between any elements in the sequence regardless of distance, allowing efficient modelling of both local and global relationships. For foundation models operating on time series data, this property is crucial, as it allows learning complex patterns and relationships within data without requiring explicit feature engineering.

Vision Transformers (ViT) [5] demonstrated that this architecture could be successfully applied beyond language to image data, by splitting images into patches and treating them as a sequence of tokens. Their impressive performance on image classification tasks highlighted the transferability of transformer architectures to other multi-dimensional data domains. Based on this success, transformers are promising candidates in other domains such as EEG, where multi-channel signals can similarly be viewed as sequences with spatial relationships.

However, transformers face a significant limitation in their quadratic computational complexity with respect to sequence length. The self-attention operation requires computing pairwise relationships between all elements, resulting in $O(n^2)$ complexity for a sequence of length n. For long EEG recordings with high sampling rates, this can become prohibitively expensive in terms of computation and memory requirements.

To reduce the total sequence length, while preserving the ability to extract local features, transformers are often paired with convolutional neural network (CNN) tokenizers [16, 17]. These tokenizers patch and embed the input signals into tokens before passing them to the transformer, effectively reducing sequence length while preserving essential fine-grained spatial information. This hybrid approach combines the local feature extraction capabilities of CNNs with the global relationship modeling of transformers, and has proven effective for processing raw physiological signals like EEG [18].

2.3.1.1. Attention

Attention mechanisms form the core of transformer architectures, enabling models to selectively focus on relevant parts of input data when generating output elements. The self-attention mechanism, introduced in the original transformer model [15], computes relationships between all positions in a sequence by projecting the input into query (Q), key (K), and value (V) representations.

For each position in the sequence, attention weights are computed as the scaled dot product between its query vector and the key vectors of all positions, followed by softmax normalization. These weights determine how much information from each position contributes to the output representation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multi-head attention extends this by running multiple attention operations in parallel with different learned projections, allowing the model to attend to information from different representation subspaces. This enables capturing diverse relationships within the data.

The primary limitation of standard self-attention is its quadratic computational complexity with respect to sequence length, as it computes interactions between all pairs of tokens. For EEG data with multiple channels and long sequences, this can quickly become computationally prohibitive.

2.3.1.2. Alternating Attention

Alternating Attention [19] addresses the computational limitations of standard self-attention for multi-channel temporal data like EEG by decomposing the attention computation into separate temporal and spatial dimensions.

The mechanism works by alternating between two types of attention across transformer layers:

- **Intra-channel (temporal) attention**: Computes attention independently within each channel across time points, with complexity $O(CN_p^2)$ where C is the number of channels and N_p is the number of patches per channel.
- **Inter-channel (spatial) attention**: Computes attention across channels at each time point, with complexity $O(C^2 N_p)$.

This approach significantly reduces memory requirements compared to standard self-attention, which has complexity $O((CN_p)^2)$.

The implementation involves reshaping the input tensor to perform attention along different dimensions in alternating layers. Odd-numbered layers perform inter-channel attention by reshaping the tensor to attend across channels, while even-numbered layers perform intra-channel attention by reshaping to attend across temporal patches within each channel.

This approach is particularly valuable for EEG processing, as it efficiently captures both spatial relationships between electrodes and temporal dynamics within each channel's signal.

2.3.2. MAMBA

The Mamba [20] model architecture represents a recent alternative to transformer-based architectures for sequence modelling. As a form of State Space Models, Mamba addresses the computational limitation of transformers, thanks to its linear computational complexity with respect to sequence length $O(n)$, while achieving similar performance. The improved efficiency enables the processing of longer continuous segments without the need for excessive windowing or downsampling, potentially preserving important temporal patterns in the signal. This makes Mamba an attractive alternative for EEG analysis, where long sequences are common and preserving temporal relationships is crucial for accurate modelling.

2.4. Self-Supervised Learning for Foundation Models

As large foundation models require vast amounts of data for training, the availability of labelled data often becomes a bottleneck in their development. Self-supervised learning (SSL) has emerged as a powerful approach to leverage unlabelled data for pretraining foundation models. Rather than relying on human-annotated labels, which are often expensive to acquire, self-supervised methods are designed to use some inherent aspects of the data itself as the supervisory signals. This makes self-supervised learning particularly valuable when labelled data is scarce but unlabelled data is plentiful.

In the context of EEG, acquiring labelled data is particularly challenging, as it requires human domain expertise and controlled experimental conditions, often making data collection and annotation prohibitively expensive and time-consuming, as well as suffering from varying expert assessments [21]. Acquisition of raw EEG data without annotation is comparatively easy and cheap, with research institutions and clinical environments generating substantial volumes of unlabelled EEG data. While such data is unsuitable for training a model on specific downstream tasks, it does share the general characteristics of EEG data and thus holds potential for models to effectively learn the data distribution and improve their generalization capability.

Self-supervised learning provides a methodology to capitalize on such more readily available unlabelled data. In this paradigm, a foundation model, consisting of a feature encoder and potentially a pretraining-specific head, is pretrained on a large amount of unlabelled data. The goal of this pretraining stage is for the encoder stage of the model to learn effective extraction of features from the data and embed these features in a meaningful latent space. For this purpose, the data, either in raw form or through some transformation, serves both as model input as well as pseudo-labels.

In the following downstream task training stage, the pretraining head of the foundation model is discarded and the feature-extracting encoder is combined with a task-specific model head and subsequently trained on the downstream task, using labelled data. If the domain shift between the pretraining and the downstream data is low, such a model can be expected to perform better than a model purely trained on the specific task using only the labelled data, especially when the volume of labelled data is significantly limited. This transfer learning approach is particularly valuable for EEG analysis, where datasets are often collected under different protocols or with different equipment, leading to significant variability in the signal characteristics.

Two prominent approaches for such self-supervised learning are masked reconstruction and contrastive learning. These methods differ in how they form and utilize pseudo-labels derived from the input data, as well as their respective model architectures.

2.4.1. Masked reconstruction

Masked reconstruction is a commonly used self-supervised learning strategy for feature extraction in images, using vision transformer models [22, 23]. While these models have proven very effective in visual tasks, the learning paradigm itself is not limited to applications in images and can be adapted to any arbitrary signal type and lends itself well to the nature of multi-channel EEG signals.

Portions of input signals are masked and used as model input, whereas the unmasked signals represent the target label that is to be reconstructed. Foundation models trained with masked reconstruction typically consist of an Encoder section, which transforms the masked input signals into a latent embedding space, and a Decoder section, which tries to reconstruct the original unmasked signal. The model thus effectively acts as a denoising autoencoder.

During training, the model learns to infer missing information from surrounding context, requiring it to develop an understanding of the underlying signal structure and relationships between different components of the data. This pretext task forces the encoder to capture meaningful representations that encode sufficient information for the decoder to perform accurate reconstruction. The quality of the reconstructions in masked regions of the input signal can serve as a measure of how well the encoder has learned to extract and embed the essential characteristics of the data.

A key advantage of masked reconstruction for EEG analysis is its ability to handle the multi-dimensional nature of the signals without requiring explicit domain knowledge about what signal characteristics are important. Instead, the model learns to prioritize features that are most useful for reconstruction, which often correlate with features relevant for downstream tasks.

2.4.2. Contrastive Learning

Contrastive learning represents an alternative training objective that operates on the principle of learning discriminative representations without requiring signal reconstruction. Unlike masked reconstruction methods, contrastive learning employs an encoder-only architecture, without the need for an additional pretraining-specific task head.

The fundamental mechanism involves creating pairs of signal examples that are either semantically similar (positive pairs) or dissimilar (negative pairs). For EEG data, these relationships can be established through various criteria: segments from the same subject, temporally adjacent recordings, or signals recorded under similar experimental conditions may constitute positive pairs, while segments not fitting the defined criteria form negative pairs. Such a criteria is especially easy to define in labelled data, where positive pairs may be formed by selecting segments from the same class label, while negative pairs are formed by selecting segments from different classes. Alternatively, a positive pair may also be formed by creating two different views from a single sample, through suitable transformations or augmentations. In this case, any two views formed from different samples constitute a negative pairing. Such a criteria is particularly useful for self-supervised training, where no class information is available.

Having defined positive and negative pairings, the training objective is to then optimize the distances between embeddings of pairs of samples. A distance loss enforces proximity of positive pairs in the latent embedding space, while negative pairings are pushed apart. This creates a latent space where semantically related signals cluster together, while unrelated signals remain separated. Such a latent space lends itself well to downstream tasks such as classification, where a classification head effectively tries to separate embeddings of different classes. Given a well-chosen clustering criterion that correlates to the downstream class distributions, the emerging decision boundaries may align well with class distinctions.

Chapter 3

Related Work

3.1. Foundation Models for EEG

Foundation models have demonstrated remarkable success in various domains, including natural language processing and computer vision, by leveraging large datasets for pre-training and enabling effective transfer learning. The application of foundation models to electroencephalography (EEG) aims to address the challenges posed by data variability and the scarcity of labelled data in EEG research. These models strive to learn generalizable representations from EEG signals that can be fine-tuned for specific downstream tasks.

Several EEG foundation models have emerged, each with its own strengths and limitations. BENDR [18] utilizes a convolutional architecture inspired by speech recognition models, focusing on learning compressed representations of raw EEG signals. BrainBERT [24] adapts the Transformer architecture, employing masked autoencoding on spectrogram patches of intracranial EEG data. LaBraM[25] introduces a learned neural tokenizer to encode EEG channel patches, which are then used to pre-train Transformer models with masked reconstruction. Similarly, EEGFormer [26] adopts a discrete representation learning algorithm using vector-quantization along with reconstruction. MAEEG [27] explores masked auto-encoders for learning EEG representations, adopting a BENDR-inspired architecture, with findings suggesting that masking larger segments improves downstream performance.

More recent models like CeReBro [19] and CBraMod [6] have addressed some limitations by developing specialized architectures for EEG. CBraMod incorporates a criss-cross transformer to model spatial and temporal dependencies separately, while CeReBro introduces alternating attention to efficiently capture both dimensions. FEMBA [7] introduces a Mamba-based architecture, establishing new efficiency benchmarks for EEG analysis through bidirectional state-space modelling.

While these models have shown promising results, many rely on complex architectures or introduce additional components like neural tokenizers, which can increase computational overhead. Furthermore, the exploration of different self-supervised

learning (SSL) tasks for EEG foundation models remains limited. Most existing models primarily focus on masked autoencoding or simple contrastive learning approaches and explore different model architectures instead.

3.2. Self-Supervised Learning Strategies

Self-supervised learning has become a prominent technique for pre-training models on unlabelled data. Two common SSL strategies are masked reconstruction and contrastive learning.

Masked reconstruction involves masking portions of the input data and training the model to reconstruct the missing information. This approach has been successfully applied in both the image and time series domains. In the image domain, models like ViT [5], ChannelViT [28], MAE [29], and SimMIM [23] have demonstrated the effectiveness of masked image modelling for learning visual representations. These approaches have directly influenced EEG foundation models such as BrainBERT and MAEEG, which adapt similar masking strategies to the unique temporal and spatial characteristics of EEG signals.

Contrastive learning, on the other hand, aims to learn representations by contrasting positive and negative examples. While contrastive learning has been explored extensively in vision and language domains, its application to EEG foundation models remains relatively limited. One notable exception is the work by Huang et al. [30], who propose a learning framework that extends masked reconstruction with a contrastive learning objective, using a momentum branch processing unmasked images. This hybrid approach hints at the potential benefits of combining multiple SSL strategies in the EEG domain, similar to JEPA frameworks.

Beyond general EEG foundation models, some studies have explored self-supervised learning for bio-signals on specific tasks. JEPA-based approaches, such as S-JEPA [31] (for EEG) and ECG-JEPA [32], have shown promise in learning representations from bio-signals by predicting masked signal portions at the hidden representation level.

3.3. Research Gap and Contribution

The review of existing literature reveals a gap in the exploration of self-supervised learning strategies for EEG foundation models. Current EEG foundation models have primarily focused on a limited set of SSL tasks, mainly masked autoencoding and basic contrastive learning. Despite the evolution from BENDR's contrastive reconstruction approach to the various masked autoencoding strategies employed by later models, the potential of other SSL strategies and their combinations remains largely unexplored.

Specifically, variations of masked reconstruction techniques, more sophisticated contrastive approaches, and hybrid methods like those combining masked reconstruction with JEPA have not been thoroughly investigated in the EEG domain. This limitation is particularly noteworthy given the success of diverse SSL strategies in other domains such as computer vision and natural language processing.

This work addresses this gap by systematically exploring different masked reconstruction tasks and contrastive learning approaches for EEG pre-training. By evaluating various SSL strategies and their combinations, this research aims to identify the most effective pre-training methods for learning robust and generalizable EEG representations.

Chapter 4

Implementation

4.1. Models

This section details the neural network architectures employed to evaluate the proposed pretraining methods. All models share a similar high-level structure, consisting of three main components: a patch embedder for tokenization, an encoder producing embeddings, and a decoder for reconstruction. During finetuning, the decoder is replaced with a linear classification layer. All models process robust-normalized EEG signals, either as full sequences or in patches. The model hyperparameters are listed in appendix B.

4.1.1. MEST with Alternating Attention

MEST (Masked EEG Sequential Transformer) is a transformer-based architecture implemented for this thesis. It is designed as a relatively straightforward architecture without novel features, making it a representative model for general EEG transformer architectures, while allowing for quick prototyping and evaluations of pretraining methods without sacrificing performance.

The patch embedder uses a convolutional approach adapted from FEMBA (see section 4.1.2), which efficiently extracts features from raw EEG signals. The transformer encoder employs Alternating Attention, a mechanism introduced in CeReBro [19], which alternates between temporal and spatial attention across layers, as explained in background section 2.3.1.2. Due to memory constraints associated with the alternating attention implementation, which combines the non-attended dimension with the batch dimension, the patch size was set to 32 samples, resulting in more manageable sequence lengths. The decoder consists of a single transformer decoder layer that operates on the flattened sequence-channel dimension, reconstructing the original input from the encoded representation.

This model was employed for the majority of the pretraining experiments to perform thorough explorations of the pretraining methods and hyperparameters. Figure 4.1 shows the general architecture of the MEST model, which is also representative of

the other transformer-based masked reconstruction models used in this thesis. The Transformer layers in this case employ alternating attention, while the convolutional patch embedder, borrowed from FEMBA, is shown in figure 4.2a.

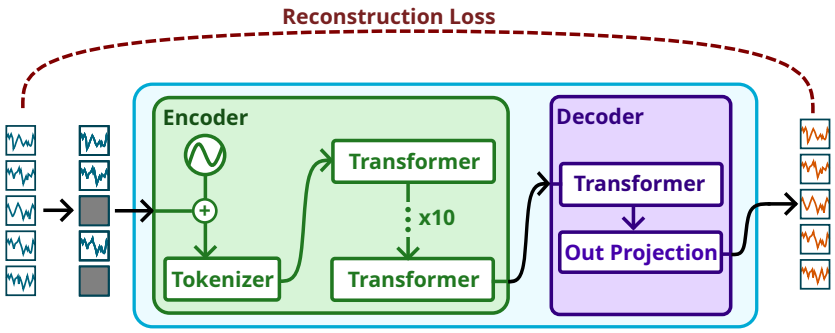
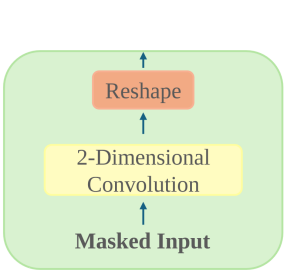


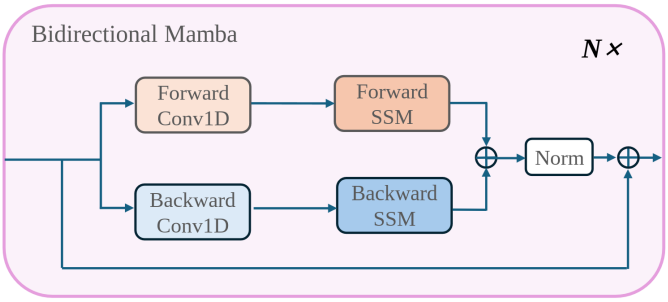
Figure 4.1.: General Encoder-Decoder architecture. The encoder processes the masked input EEG data, while the decoder reconstructs the original signal. The encoder uses alternating attention to capture temporal and spatial relationships, while the decoder employs a conventional transformer layer to generate the output.

4.1.2. FEMBA

FEMBA [7] was introduced as a state-space model alternative to transformer-based architectures for EEG processing, and is based on Mamba as introduced in background section 2.3.2. It leverages the Mamba architecture [20] to achieve efficient long-sequence processing with linear scaling properties.



(a) FEMBA convolutional Tokenizer.



(b) FEMBA Encoder.

Figure 4.2.: FEMBA architecture components. Figures adapted from Tegon et al. [7].

The model introduces the same convolutional patch embedder also used in the MEST implementation. Instead of attention-based transformer layers, the encoder utilizes

state-space Mamba blocks that enable processing of longer sequences with reduced computational requirements. As in MEST, the decoder consists of a single transformer decoder layer operating on the flattened sequence-channel dimension.

For the evaluation experiments, the FEMBA-tiny configuration is used, as it was shown to achieve similar performance as the larger model sizes while being faster to train. According to the original paper, this smaller variant achieves comparable results to larger configurations while requiring significantly less computational resources. Thanks to quick training times, this model has also been used for a large portion of the pretraining evaluations.

4.1.3. LaBraM

LaBraM is an established EEG foundation model developed by Jiang et al. [25]. It employs a transformer-based architecture with a two-stage pretraining strategy: first training a vector-quantization tokenizer with a codebook, followed by masked reconstruction with a patch size of 200. The availability of publicly released pretrained weights makes LaBraM particularly valuable as a benchmark. It allows for direct comparison between our best-performing pretraining methods and an established baseline by fine-tuning both on the same datasets. The model architecture follows the transformer paradigm with a patch embedder and transformer encoder, completed by a single-layer transformer decoder operating on the flattened sequence-channel dimension.

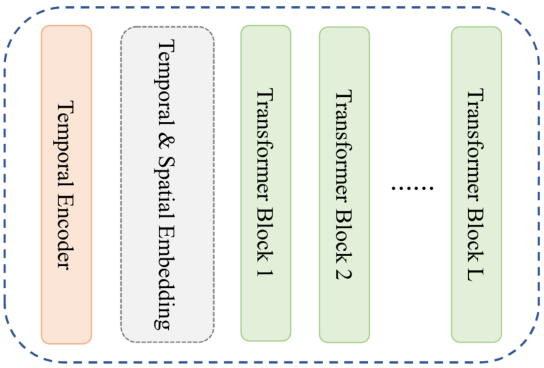


Figure 4.3.: LaBraM Encoder component, also containing Tokenizer, here called Temporal Encoder. For the full reconstruction model, the Encoder is paired with a similar Decoder Head as used in MEST and FEMBA.
Figure adapted from Jiang et al. [25].

4.1.4. MEST with sparse attention

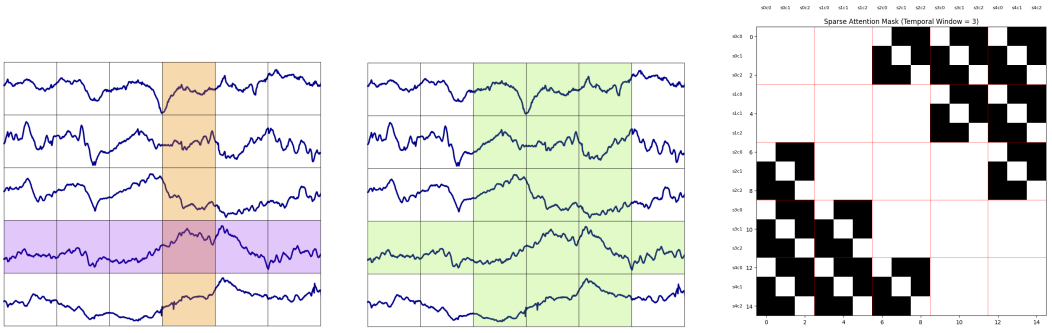
This variant investigates whether significant information exists in cross-channel, cross-temporal correlations in EEG signals. While alternating attention is capable of capturing intra-channel and intra-temporal patterns, it may miss interactions where brain activity at one location propagates to different electrodes with temporal delay.

To capture these relationships while maintaining reasonable memory requirements, this MEST variant implements a sparse attention mechanism over a flattened sequence-channel dimension in the encoder. The attention mask is specifically designed to allow:

- Full temporal attention within each channel
- Cross-channel attention within a limited temporal neighbourhood

The temporal neighbourhood window spans 7 patches, resulting in approximately 20% of all possible token pairs participating in attention computations. This approach offers a compromise between computational efficiency and the model's ability to capture complex spatio-temporal relationships in EEG data.

Due to computational constraints further detailed in section 5.1.1.1, this model could not be fully trained and has not been further evaluated.



(a) Alternating attention (b) Sparse Attention (c) Sparse Attention mask

Figure 4.4.: Attention mechanisms in MEST transformer models: (a) Alternating attention separates temporal and spatial dimensions, (b) Sparse attention allows cross-dimensional interactions within a local neighborhood, and (c) Visualization of the sparse attention mask on a 3-channel signal over 5 temporal steps, temporal attention window size 3, showing which token pairs participate in attention computations in white. Larger squares in red represent temporal patches, subdivisions represent 3 channels per signal.

4.2. Masked Reconstruction Pretraining Methods

Masked reconstruction forms the basis of the self-supervised learning methods discussed in this section. The input EEG data is first divided into 80 patches of size 16 (equivalent to 62.5ms at 256Hz) along the temporal dimension. These patches are then used as input for the pretraining methods, with the original data serving as the reconstruction target.

The general workflow for all masked reconstruction methods follows these steps:

1. Patch the input EEG data into fixed-size segments

2. Replace selected patches with fixed mask token according to masking strategy: $x_m = \mathcal{M}(x)$
3. For model-specific processing, masked data is reassembled and re-patched according to each model's requirements (MEST: 32 samples, LaBraM: 200 samples), while FEMBA processes the unpatched input directly
4. The model reconstructs the original signal from the masked input: $\hat{x} = f(x_m)$
5. Reconstruction quality is evaluated using smooth L1 loss between the model output $\hat{x}$ and the reconstruction target x: $\mathcal{L}_{\text{reconstruction}} = \text{SmoothL1}(x, \hat{x})$

The Smooth-L1 loss [33] is preferred over standard MSE or L1 loss as it combines the advantages of both: behaving like L2 (MSE) loss for small errors to provide stable gradients near the minimum, while behaving like L1 loss for larger errors to reduce the impact of outliers. This is particularly beneficial for the TUEG EEG data, which contains some significant outliers, see section 4.5. The loss function is defined as:

$$\text{SmoothL1}(x, \hat{x}) = \begin{cases} 0.5 \cdot (x - \hat{x})^2, & \text{if } |x - \hat{x}| < 1 \\ |x - \hat{x}| - 0.5, & \text{otherwise} \end{cases} \tag{4.1}$$

where x represents the target signal and $\hat{x}$ is the model's prediction.

Unmasked patches are also reconstructed and included in the loss calculation with a weight of 0.1 to ensure that the model learns to reconstruct the entire signal, not just the masked regions.

The masked reconstruction strategies generally use a mask ratio of 0.5. Different mask ratios were evaluated in earlier experimentations, ranging from 0.3 to 0.9. Visual inspection of reconstructions and early finetuning results on downstream tasks indicated that ratios between 0.5 and 0.6 provided the best balance between reconstruction difficulty and amount of visible signal data. This empirical finding aligns with the established literature, particularly the ablation studies conducted by Jiang et al. [25] for LaBraM. Consequently, a mask ratio of 0.5 was adopted for all final evaluations.

4.2.1. Random Patch Masking

Random patch masking follows the masking procedure established in image-based SimMIM [34] and adapted to multi-channel EEG data. In this approach, patches are randomly selected for masking with the mask ratio of 0.5.

A key implementation detail is that patches are masked independently across channels. This independence encourages the model to learn inter-channel relationships, as information absent in one channel must be inferred from other channels where the corresponding temporal segment remains visible. The approach leverages spatial redundancy inherent in EEG signals, where neural activity is often reflected across multiple electrodes with varying amplitude and phase.

The random distribution of masks creates diverse reconstruction challenges during training, forcing the model to develop robust representations that capture both local

and global signal characteristics. This method serves as our baseline masking strategy against which other techniques are compared.

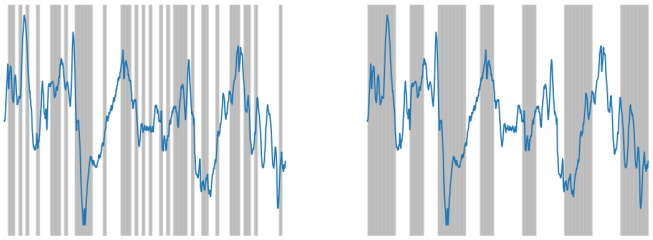
4.2.2. Clustered Random Patch Masking

A limitation of standard random masking is that spatially isolated masked patches may be too easily reconstructable through simple interpolation from adjacent patches, potentially limiting the model's need to learn deeper contextual representations. Clustered random patch masking addresses this limitation by ensuring that masked regions form contiguous segments of a minimum width while maintaining the same overall masking ratio of 0.5.

The implementation uses a resolution-reduction approach:

1. First, a lower-resolution mask is generated by dividing the temporal dimension by a cluster size factor
2. Random masking is applied at this reduced resolution, where each mask element corresponds to multiple consecutive patches
3. The mask is then upsampled back to correct size by cluster size factor

The implementation maintains channel independence, applying the same clustering approach separately to each channel. This ensures that reconstruction tasks require understanding both local temporal context and cross-channel relationships. The clustered approach creates a more challenging reconstruction task that cannot be solved through simple interpolation between adjacent time points, encouraging the model to learn broader contextual patterns within the data.



(a) Random (b) Random Clustered

Figure 4.5.: Comparison of masking strategies: (a) Random masking and (b) Random Clustered masking, applied to single channel signal. Masked patches in grey.

4.2.3. Lowpass-filtered Reconstruction Target

EEG signals are characterized by a complex mixture of frequency components, as described in background section 2.1, with research indicating that lower frequencies (below 50Hz) contain most of the task-relevant information for many cognitive processes [35, 36].

Higher frequencies often contain a significant proportion of noise and artifacts that may not contribute meaningfully to downstream tasks.

This method exploits this frequency characteristic by using a lowpass-filtered version of the original data as the reconstruction target while maintaining the standard masked data as input. Specifically, the implementation applies a 5th-order Biquad lowpass filter with a cutoff frequency of 50Hz to the original signal to create the target for reconstruction. During training, the loss is calculated between the model output and the lowpass-filtered target rather than the original signal.

The reconstruction objective thus becomes learning a lowpass-filtered and ideally de-noised representation of the EEG signal, rather than reproducing the original signal precisely. This approach effectively incorporates a form of signal denoising into the pretraining process, potentially yielding more robust representations for downstream tasks.

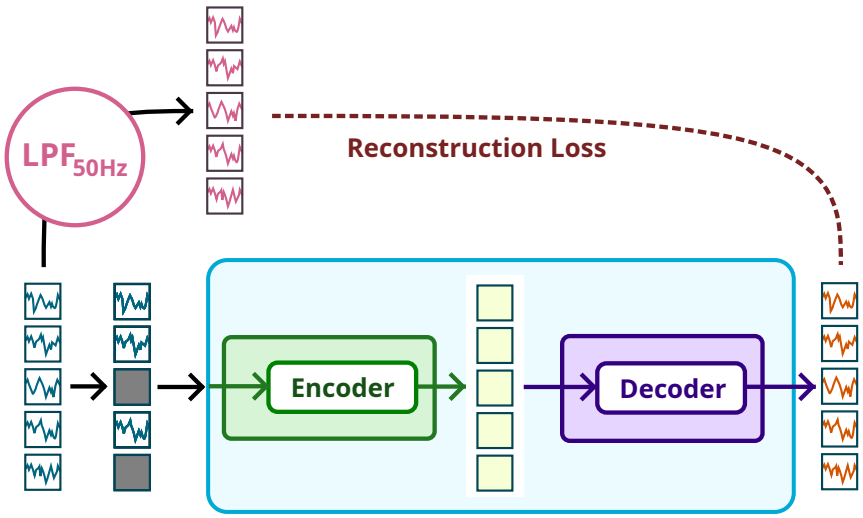


Figure 4.6.: *Lowpass-Filtered Reconstruction Target* pretraining scheme.

This method may be applied in combination with any of the masked reconstruction strategies.

4.2.4. JEPA

Joint Embedding Predictive Architecture (JEPA) is a self-supervised learning technique originally proposed by Assran et al. [37] for computer vision and adapted for EEG data in recent works [31, 32]. Our implementation differs from these prior studies by using JEPA as a complementary objective alongside masked reconstruction rather than as a stand-alone pretraining method. As such it is a similar framework as proposed

by Huang et al. [30] for vision transformers, although with a different loss function compared to their contrastive loss.

The motivation for incorporating JEPA stems from two fundamental limitations observed with pure masked reconstruction approaches:

- **Noise sensitivity:** Reconstruction-based methods aim to replicate the input signal precisely, including any noise or artifacts present in the original data. This can limit the model's ability to distinguish between meaningful neural activity and irrelevant noise.
- **Misalignment of loss and objective:** The reconstruction loss in the input space may not be a good descriptor for the model's ability to extract meaningful features, and thus can be a misguided objective to optimize. To use a visual example: suppose a visual model may be equally adept at extracting the features designating a blue sky or a hand. If both features are equally relevant for downstream tasks, the ideal loss would guide training to improve extraction of both equally. However, the hand is inherently more complex to reconstruct due to its intricate details, resulting in higher reconstruction loss regardless of feature importance. However, the higher complexity of the hand is not necessarily indicative of better or more relevant features. When operating in the input space, the loss is unavoidably biased by the reconstruction difficulty, which may lead to disproportionate optimization focus on complex patterns that may not necessarily contain more relevant information.

The JEPA implementation employs a secondary teacher model to calculate the JEPA loss alongside a student model, trained via masked reconstruction.

1. The **student network** follows the standard encoder-decoder architecture used in masked reconstruction, receiving masked input and being trained through backpropagation.
2. The **teacher network** consists only of an encoder (without a decoder) and is updated as an exponential moving average (EMA) of the student's encoder weights. The teacher receives and processes the complete unmasked signal as input.
3. The total JEPA loss is calculated as

   $$\mathcal{L}_{\text{JEPA}} = 1 - \text{cosine\_similarity}(z_t, z_s)$$

   , where:
   - z_t is the embedding produced by the teacher encoder from unmasked input
   - z_s is the embedding produced by the student network at an intermediate stage of decoding
4. The final training objective combines the JEPA loss with the traditional masked reconstruction loss:

   $$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{reconstruction}} + \lambda \cdot \mathcal{L}_{\text{JEPA}}$$

   , where λ controls the contribution of the JEPA component.

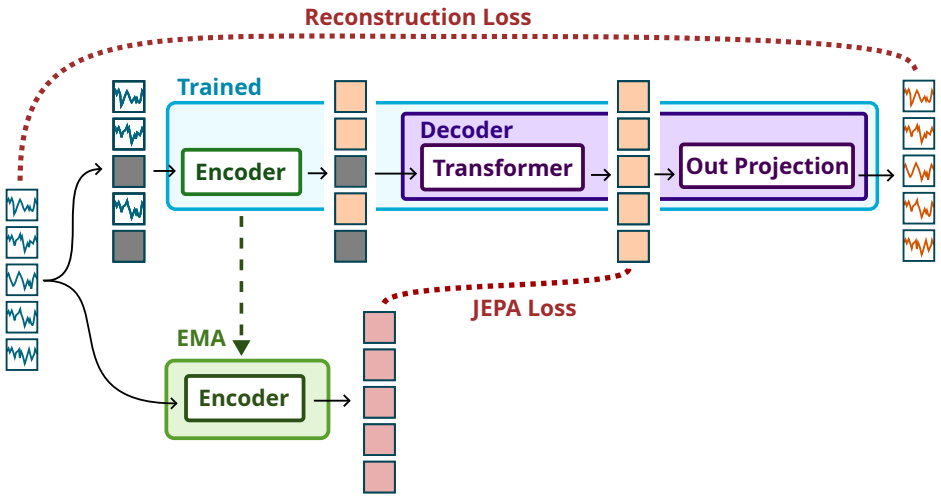


Figure 4.7.: Joint Embedding Predictive Architecture (JEPA) implementation. The student network (top) processes masked EEG data, while the EMA teacher network (bottom) processes unmasked data. The JEPA loss is calculated between the teacher's output embedding and an intermediate embedding from the student's decoder.

This approach effectively creates a dual learning objective: the reconstruction loss trains the model to predict raw signal values, while the JEPA loss encourages learning representations that capture the essential information needed to create embeddings matching those derived from complete data. By operating in the latent space rather than the raw signal space, JEPA helps focus learning on meaningful features rather than precise signal reconstruction.

An important implementation detail is the selection of which student embedding to compare with the teacher embedding. By taking the student embedding part-way through the decoder (after the transformer decoder layer but before output projection), we allow the decoder to perform "latent in-painting" of masked portions before comparison. This creates a more appropriate comparison point with the teacher's complete-data embedding than using the student-encoder output directly.

4.2.5. Further MAE Experiments

Beyond the primary masked reconstruction approaches described above, several exploratory experiments were conducted with alternative pretraining strategies. These approaches, while theoretically promising, demonstrated limited effectiveness in early experiments and were therefore not included in the final evaluation. The following sections describe these approaches, the rationale behind them, and the specific challenges that prevented their further exploration.

4.2.5.1. Forecasting Masking

In addition to the random and clustered masking strategies, a forecasting masking approach was also explored, where the rear temporal portion of the input sequence is masked. Unlike random masking, which distributes masked regions throughout the signal, forecasting masking creates a prediction task that requires reconstructing future signal values based solely on past observations.

This paradigm presents a more challenging reconstruction task for two key reasons:

1. **Continuous masking with one-sided context:** The masked region forms a single long continuous section with unmasked signal available only at the temporally older side of the masked region, eliminating most of the available within-channel surrounding contextual information.

2. **No cross-channel context:** Since all channels are masked in the same temporal region, no contextual information from adjacent channels in the masked region is available.

The underlying hypothesis was that this increased difficulty would force the model to develop a deeper understanding of temporal dynamics and signal evolution patterns rather than relying on simple interpolation or cross-channel inference.

However, despite extensive experimentation, forecasting masking consistently failed to produce meaningful representations. In all configurations, models trained with this approach converged to trivial solutions, primarily outputting constant values approximating the signal mean in the masked regions.

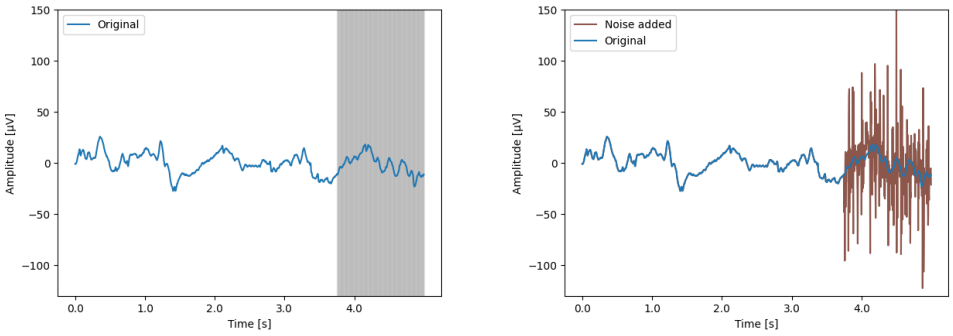
Several variations to improve performance were explored:

- **Curriculum Learning**: Gradually increasing the mask ratio over training to allow the model to learn from easier reconstruction tasks first, see also section 4.2.5.3.

- **Noise-based Masking**: Rather than complete replacement with mask tokens, adding increasing amounts of Gaussian noise to the masked region, with noise magnitude increasing either over training epochs or with distance from the last observed value.

- **Loss weighting strategies**: Applying variable weights to the reconstruction loss based on temporal distance from the last observed value, both with increasing and decreasing weighting.

Despite these efforts, all forecasting configurations resulted in the model producing trivial reconstructions approximating the mean signal value (see figure 4.9a). This suggests that predicting future EEG values without any surrounding context may be inherently too difficult for the current model architectures and training objectives.



(a) Forecasting mask (b) Forecasting noise mask

Figure 4.8.: Comparison of masking strategies: (a) Forecasting masking, (b) Forecasting noise masking, applied to single channel signal. Masked patches in grey.

4.2.5.2. Spectral Reconstruction Loss

In addition to standard reconstruction in the time domain, incorporating frequency-domain information through a spectral reconstruction loss was explored. This approach calculated an additional loss in the frequency domain, potentially capturing important spectral characteristics of EEG signals that time-domain reconstruction might miss.

The implementation involved transforming both the reconstruction target and the model output to the frequency domain using Fast Fourier Transform (FFT). The spectral loss was then calculated on the absolute magnitudes of the frequency components:

$$\mathcal{L}_{\text{spectral}} = \mathcal{L}(|\text{FFT}(x)|, |\text{FFT}(\hat{x})|)$$

where x represents the original signal, $\hat{x}$ is the reconstructed signal, and $\mathcal{L}$ is the same smooth L1 loss used for time-domain reconstruction. Note that unlike in time-domain reconstruction, the spectral loss is calculated with equal weighting over the whole signal, without special significance given to the masked regions.

The primary motivation for this approach was addressing the observation that model reconstructions were particularly deficient in high-frequency components. Visual analysis of reconstructions showed that models tended to capture low-frequency trends well but struggled with higher-frequency details. By explicitly including spectral characteristics in the loss function, we aimed to improve reconstruction across the full frequency spectrum. Specifically for the case of forecasting masking, the hope was that this approach would encourage some less conservative predictions.

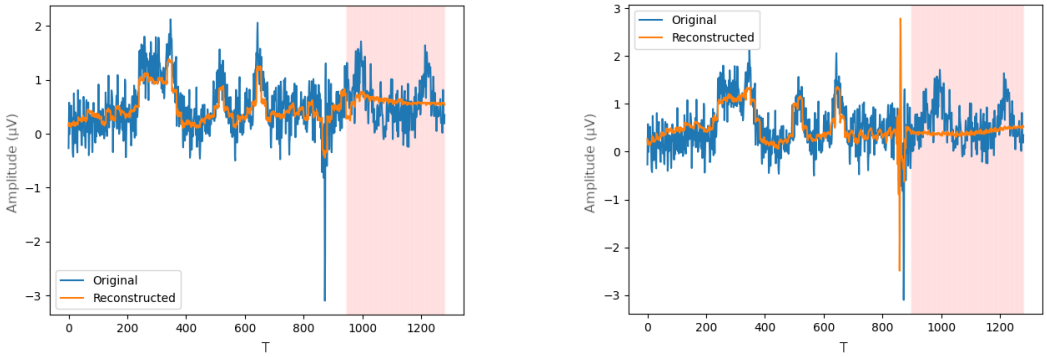
The combined loss function was defined as:

$$\mathcal{L}\text{total} = \mathcal{L}\text{time} + \lambda \cdot \mathcal{L}_{\text{spectral}}$$

where λ controlled the contribution of the spectral component.

While this approach did produce visually somewhat improved reconstructions with better high-frequency content (see figure 4.9), downstream task performance did not

improve. This suggests that the additional high-frequency components captured were likely above the relevant EEG frequency ranges (typically below 50Hz) and may have represented noise rather than meaningful signal.



(a) Forcasting Masked Reconstruction

(b) Forecasting Masked Reconstruction with Spectral Loss

Figure 4.9.: Spectral Loss improves reconstruction of high frequency spike, but not reconstruction quality in masked region

Some practical challenges also emerged during implementation:

- Computing FFT on long sequences was computationally expensive, requiring chunk-wise processing and increasing training time significantly.
- Balancing the weights between time and spectral losses required careful tuning.

Due to the lack of improvement in downstream performance and the additional complexity, this approach was ultimately not pursued further.

4.2.5.3. Curriculum Learning for Mask Ratio

To better guide the model through the training process and potentially achieve higher maximal mask ratios with good reconstruction quality, a Curriculum Learning approach for the mask ratio parameter was implemented. Curriculum learning is a training paradigm based on gradually increasing the task difficulty over time, allowing models to learn from simpler examples before progressing to more complex ones.

The implementation dynamically adjusted the mask ratio based on validation performance. The mask ratio was determined using an inverse linear map of the validation loss. This approach created a challenge, however: changing the masking ratio directly influenced the validation loss itself, creating an undesired feedback loop that could lead to oscillating behaviour.

To address this issue, a one-way progression rule was used: the mask ratio could only be increased, never decreased. The next curriculum step would then only be taken

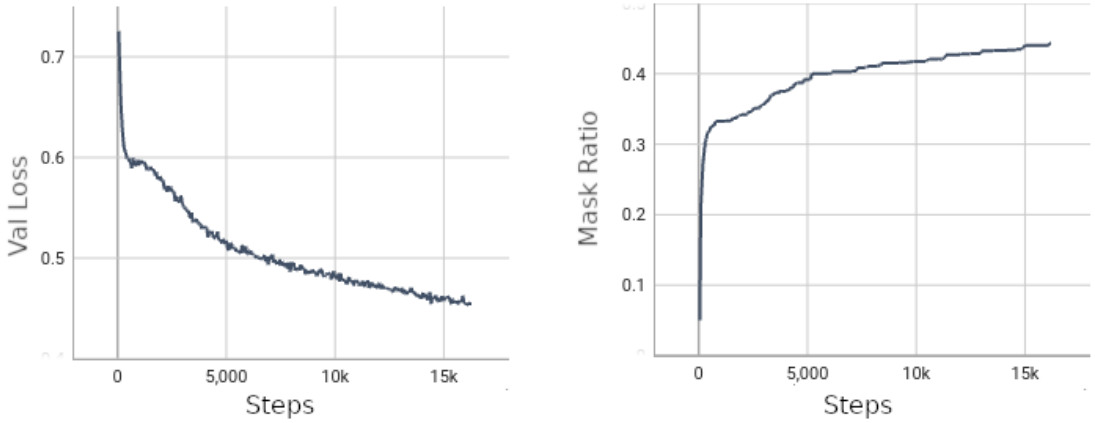
after the validation loss had decreased further following a previous mask ratio increase, indicating that the model had successfully adapted to the current difficulty level.

Besides this measure, the linear map from loss to mask ratio still required manual tuning of several parameters, such as minimum and maximum mask ratio, as well as the expected minimum validation loss. The best values for these parameters were often unique to each model and training setup.

The curriculum learning mask ratio is defined as:

$$\begin{aligned} r'(t) &= (r_{max} - r_{min}) \cdot \frac{L_0 - L(t)}{L_0 - L_T} + r_{min} \\ r(t) &= \min(r_{max}, \max(r(t-1), r'(t))) \end{aligned} \tag{4.2}$$

, with L_0 being the initial validation loss, L_T the target validation loss and (r_{min}, r_{max}) being minimum and maximum mask ratios. Figure 4.10 shows the development of validation loss and mask ratio over a training run.



(a) Validation Loss using Mask Ratio Curriculum Learning over training steps

(b) Mask Ratio using Curriculum Learning over training steps

Figure 4.10.: Development of mask ratio, ratio approaches $r_{max} = 0.5$ as validation loss nears $L_T = 0.3$.

Even with a well-tuned setup, this method did not demonstrate improved reconstruction capability compared to fixed mask ratio training. In early experiments, models trained with curriculum learning achieved similar downstream task performance as those trained with an optimal fixed mask ratio. This suggests that the benefits of curriculum learning might be limited in this specific masking-based pretraining context for EEG data. Due to the limited benefit and the additional requirement for manual parameter tuning, this approach was not further pursued.

4.3. Contrastive Learning Pretraining Methods

As explained in background section 2.4.2, contrastive learning methods aim to learn representations by contrasting positive and negative sample pairs. In the self-supervised context, where no labels are available to define positive pairs, we create these pairs through augmentations of the same underlying EEG signal. The fundamental principle is that different views of the same signal sample should have similar representations (positive pairs), while views from different samples should have distinct representations (negative pairs).

The implementation follows a standard contrastive learning framework using the InfoNCE loss [38]:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(z_i, z_j)/\tau)} \tag{4.3}$$

where z_i and z_i^+ represent embeddings of two distinct views from the same signal sample, $\text{sim}(\cdot,\cdot)$ is the cosine similarity function, τ is a temperature parameter controlling the concentration of the distribution, and the denominator sums over all other samples in the batch (including one positive and $N-1$ negatives).

Previous research by Li et al. [39] demonstrated the potential of contrastive learning for EEG data, but utilized augmentations such as stretching, squeezing, cutting and reassembling the signals. Such time-domain augmentations were shown to be suboptimal by Yang et al. [40] and Rommel et al. [41]. Furthermore, Rommel at al.'s work indicated that different augmentations may be better suited to different downstream tasks, which motivates the exploration of multiple augmentation strategies.

For all contrastive methods shown, the model architecture is modified by removing the decoder and operating on the embeddings produced by the encoder. Depending on the specific model, the encoder embeddings are averaged over the sequence dimension. The contrastive loss is then calculated on these embeddings.

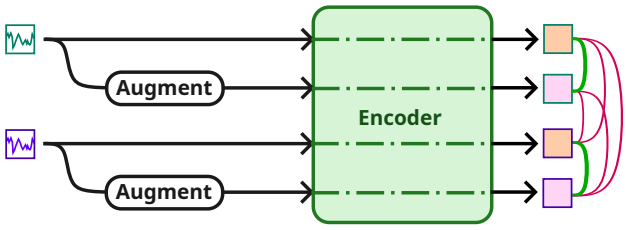


Figure 4.11.: Scheme of Contrastive learning approach for EEG signals. The Encoder processes two distinct views of each sample. Green lines indicate positive pairings (embeddings from the same original sample) that should be pulled together in the embedding space. Red lines indicate negative pairings (embeddings from different samples) that should be pushed apart. In practice, the model is trained on a batch of N samples, with each sample paired with one positive and $2N - 2$ negative samples.

4.3.1. Frequency-domain Augmentations

EEG signals contain significant information in their spectral characteristics, with distinct frequency bands corresponding to different neural activities. To learn meaningful spectral features, the frequency-domain augmentation strategy combines three complementary techniques, as presented in [41]:

- **FTSurrogate Augmentation** [42]: This technique adds random phase to the Fast Fourier Transform (FFT) of all channels while preserving the magnitude spectrum. The motivation for this augmentation is that for EEG signals, the magnitude of frequency components often carries more task-relevant information than the exact phase relationships. While the magnitude characteristics are preserved, the time-domain representation of the signal is entirely different, as is shown in figure 4.12. The transformation follows:

$$\mathcal{F}[\text{FTSurrogate}(X)](f) = \mathcal{F}[X](f)e^{i\Delta\phi(f)}$$

  where $\mathcal{F}$ denotes the Fourier transform, X is the original signal, and $\Delta\phi(f)$ is a random phase shift drawn from a uniform distribution over $[0, 2\pi]$.

- **FreqShift Augmentation** [43]: This augmentation randomly shifts the frequency components of all channels, simulating the inter-subject variability in common frequency expressions observed in EEG data. The implementation uses the Hilbert transform to perform the frequency shift on the complex analytic signal:

$$X_a(t) = X(t) + j\mathcal{H}[X](t) \tag{4.4}$$

  where $X(t)$ is the original signal and $\mathcal{H}[X](t)$ denotes the Hilbert transform. The frequency shift is then applied as:

$$\text{FrequencyShift}[X](t) = \text{Re}\left(X_a(t) \cdot e^{2\pi i \Delta f \cdot t}\right) \tag{4.5}$$

where the shift value Δf is randomly sampled from a uniform distribution over $[-\Delta f_{\max}, +\Delta f_{\max}]$, with $\Delta f_{\max} = 3$.

- **Gaussian Noise Augmentation**: Simple additive Gaussian noise (zero mean, standard deviation 0.1) is applied to all channels, enhancing model robustness against recording noise and artifacts.

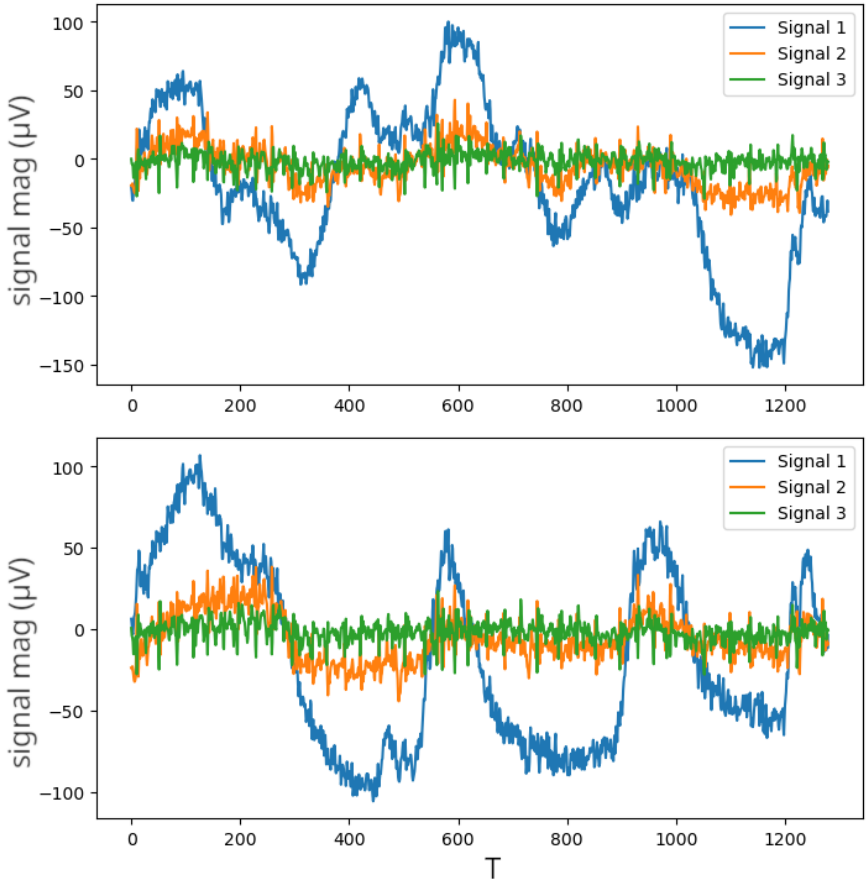


Figure 4.12.: Original signals (top), phase shifted using FTSurrogate (bottom).

These augmentations were selected based on Rommel et al.'s comprehensive analysis [41], which demonstrated their effectiveness for EEG data. For the pretraining, the positive pair is defined as the original signal and its augmented version after applying the combination of these frequency-domain transformations.

4.3.2. Topological Channel-mixing Augmentation

The topological channel-mixing augmentation exploits the spatial arrangement of EEG electrodes, leveraging the fact that neighbouring channels often record similar neural activity with slight variations. This novel approach creates augmented views by recombining information from topologically adjacent channels, ideally preserving the information content while altering the signal's time-domain expression.

The implementation consists of three components:

- **Channel Mixing**: For each channel, a weighted average is computed of its topological neighbours (excluding the channel itself). The mixing is determined by:

$$x'_i = \sum_{j \in \mathcal{N}_{3\setminus i}} w_{ij} x_j \tag{4.6}$$

  where x'_i is the mixed signal for channel i, $\mathcal{N}_{3\setminus i}$ represents the triplet of neighbouring channels (not including channel i itself), w_{ij} are randomly generated weights normalized to sum to 1, and x_j is the signal from neighbour channel j. The neighbour relationships are defined based on the physical locations of electrodes on the scalp, a neighbour set consists of 3 topologically adjacent channels (for neighbour definitions, see table B.4 in appendix B).

- **Random Scale**: A channel-wise random scaling factor drawn from uniform distribution over $[0.5, 1.5]$ is applied to simulate amplitude variations while preserving relative patterns.

- **Random Bias**: A small random offset drawn from uniform distribution over $[-1, 1]$ is added to each channel, simulating baseline shifts that commonly occur in EEG recordings.

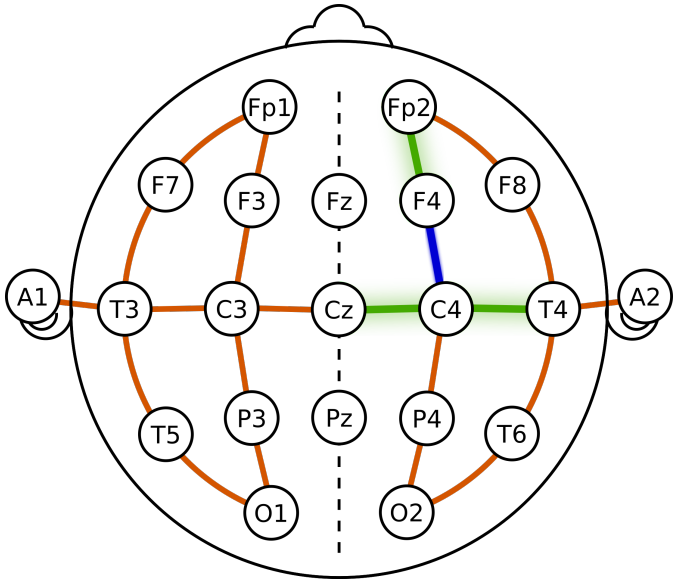


Figure 4.13.: Electrode layout in bipolar 10-20 system. Neighborhood example: channel *F4-C4* marked in blue, with its neighbor channels in green. Image Source [44].

To the best of our knowledge, this topological mixing approach represents a novel augmentation strategy for EEG data not previously explored in the literature. The positive pairs for contrastive learning consist of the original signal and its topologically mixed counterpart.

4.3.3. Masking-based Augmentation

Unlike the previous augmentation methods that transform the entire signal, the masking-based approach creates two distinct views by selectively revealing different portions of the same signal.

The motivation for this approach stems from the work Myna for musical representation learning by Yonay et al. [45]. They demonstrated that using masked views for contrastive learning rather than masked reconstruction offers significant advantages: the model can focus on extracting meaningful features (like pitch and rhythm) rather than reconstructing exact waveform details. For EEG data, this concept may also be relevant, as the model can potentially learn to identify cognitive states and neural patterns without being constrained by the need to reconstruct precise signal amplitudes.

The implementation creates two non-overlapping binary masks applied to the original signal:

- Two random positive masks M_1 and M_2 are generated, defining visible sections, while ensuring there is no overlap between them, see figure 4.14.
- Each mask reveals approximately 38% of the original signal.

- Hidden signal regions are replaced by a fixed mask token, analogous to the masked reconstruction implementation.
- The same pair of masks is applied to all samples within a batch to prevent the model from learning to match based on mask patterns rather than signal content

The resulting views are defined as:

$$v_1 = x \odot M_1 \tag{4.7}$$
$$v_2 = x \odot M_2 \tag{4.8}$$

where $\odot$ represents element-wise multiplication.

This masking-based approach differs fundamentally from the masked reconstruction method in section 4.2. While that approach aims to predict the content of masked regions, this contrastive method focuses on learning representations that are consistent across different observed portions of the same signal.

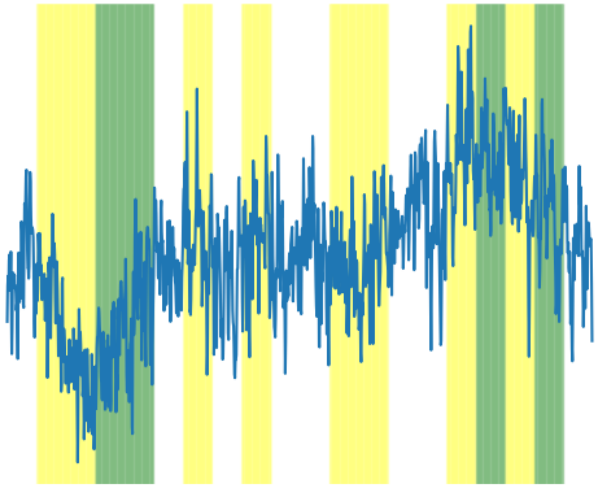


Figure 4.14.: Positive (visible) masks M_1 and M_2 in yellow and green, revealing separate signal portions used for *Masking Augmenation* pair formation.

In this augmentation settings, the original signal is not part of the positive pair, instead it is formed by the two masked views of the same signal. It this sense it differs from the previous two contrastive methods, which contrast between original and augmented view.

4.4. Finetuning

After pretraining, the models were evaluated by finetuning on downstream classification tasks. The finetuning process replaces the decoder with a lightweight classification

head while keeping the pretrained encoder architecture intact. This approach allows the models to leverage the representations learned during pretraining for specific EEG classification tasks.

The classification head consists of average pooling the embeddings over the sequence dimension and a linear projection layer to map to the appropriate number of output classes. Figure 4.15 illustrates the modified model architecture for finetuning.

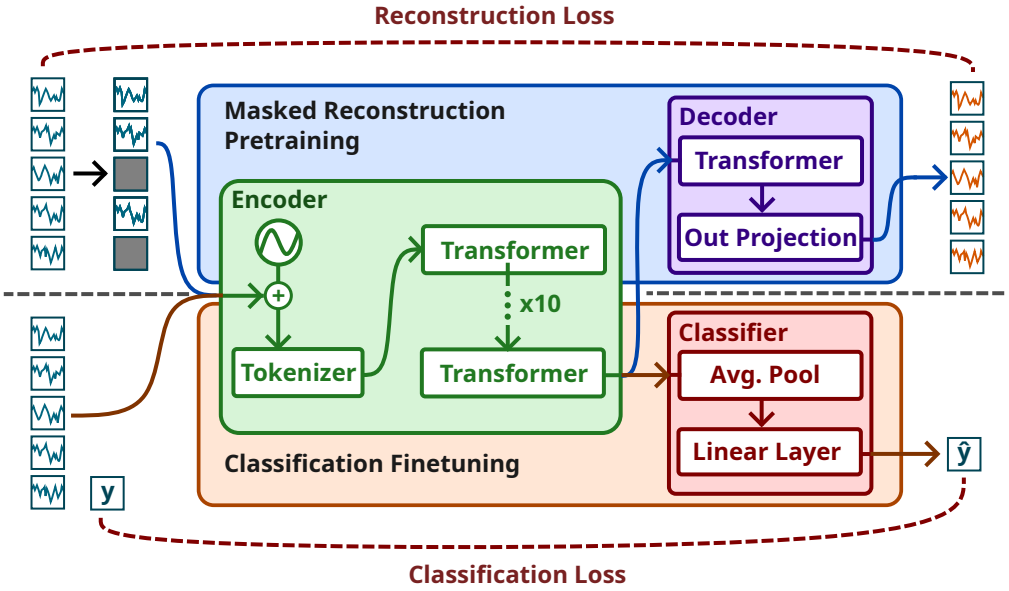


Figure 4.15.: Model architecture showing adaptation for finetuning with classification head. The pretrained encoder processes EEG data while the lightweight classification head replaces the reconstruction decoder.

Finetuning was performed on the three EEG classification datasets TUAB, TUEG and TUSL, as described in background section 2.2.2.

4.4.1. Training Methodology

The finetuning process differed in some aspects between the models and the downstream datasets. For TUAB with balanced classes cross-entropy loss was used, while weighted Focal Loss (section 4.4.1.2) was applied for the imbalanced TUAR and TUSL datasets to address class distribution issues. Class weights were set inversely proportional to class frequency. Significant overfitting was observed during initial experiments. To mitigate this, layer-wise learning rate decay (section 4.4.1.1) was employed alongside several common regularization techniques such as dropout and weight-decay (see table B.5 for details). The finetuning furthermore employed Stochastic-Weight-Averaging [46] to improve generalization. Table B.5 in appendix B details the hyperparameters used in the finetuning.

Model performance was evaluated using the following metrics:

- **AUROC** (Area Under Curve - Receiver Operating Characteristic): Measures the model's ability to discriminate between classes
- **AUPRC** (Area Under Precision-Recall Curve): Particularly informative for imbalanced datasets like TUAR and TUSL

Macro-averaged metrics were computed across the classes.

4.4.1.1. Layer-wise Learning Rate Decay

A critical component to address overfitting was implementing a layer-wise learning rate decay (LwLRD). Learning rates were scaled by a decay factor for deeper layers of the pretrained encoder, with higher learning rates for layers closer to the output. This preserved the encoders well-generalized feature extracting capabilities of the lower layers, while allowing task-specific adaptation of later layers, mainly the classifier.

4.4.1.2. Focal Loss

For the imbalanced TUAR and TUSL datasets, Focal Loss [47] was employed to address class imbalance:

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t)$$

where p_t is the model's estimated probability for the correct class, α_t is a class-balancing weight, and γ is the focusing parameter. Focal Loss offers two key advantages over standard cross-entropy:

- Down-weighting easy examples: The $(1-p_t)^\gamma$ term reduces the contribution from well-classified examples, allowing the model to focus on harder, often minority-class examples that would otherwise be overwhelmed by the abundant majority class samples.
- Class balancing: The α_t factor explicitly weights classes inversely to their frequency, preventing the model from achieving high accuracy by simply predicting the majority class.

This approach proved particularly valuable for TUSL, where significant class imbalance exists, ensuring all classes receive appropriate attention during training regardless of their representation in the dataset.

4.5. Data Characteristics

4.5.1. Pretrain Data Characteristics

Understanding the characteristics of the EEG data proved crucial for developing effective pretraining strategies. Initial experiments were plagued by training instabilities due to *NaN* values, prompting a detailed investigation into the data distribution.

Analysis of the TUEG dataset revealed significant outliers in the signal amplitude distribution. While the majority of the data exhibited reasonable amplitude ranges, a small percentage of values were extremely large, potentially causing gradient explosion and *NaN* values during training. Specifically, we found that the inner 96% of mean signal values fell within a relatively narrow range (-20.16 to 19.96), while the remaining 4% contained values with substantially higher magnitudes, with absolute maximal values exceeding 4000. The distribution of the signal mean values is shown in figure 4.16.

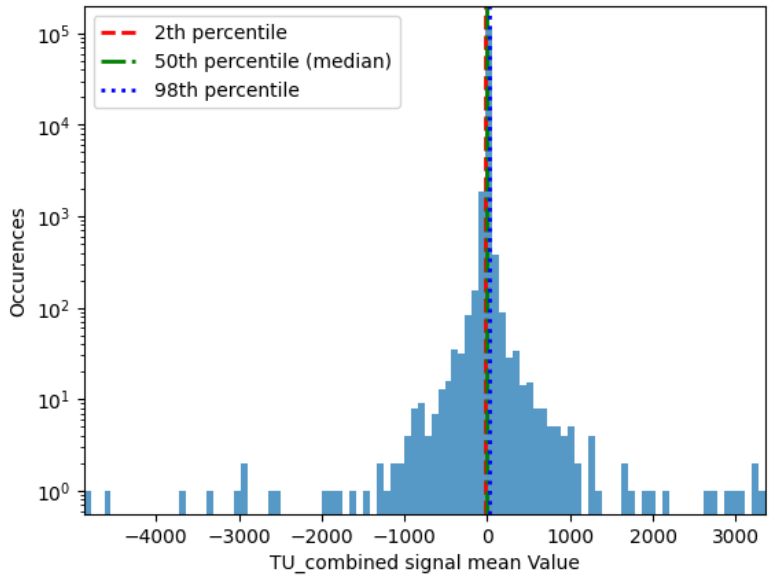


Figure 4.16.: Distribution histogram of mean signal values across the TUEG dataset, y log axis showing number of occurences. The 2nd and 98th percentiles, as used by robust normalization, are marked.

Despite the presence of these extreme values, the deliberate decision not to remove outliers was made for two reasons:

- **Maintaining dataset integrity**: To enable fair comparison with other works that use the full TUEG pretraining corpus
- **Enhancing model robustness**: Experiments conducted after resolving initial stability issues indicated that outlier removal actually decreased downstream performance, likely by reducing the model's exposure to challenging samples

Instead of outlier removal, two measure were taken to address the data distribution challenges:

- **Smooth-L1 Loss**: As described in section 4.2, this loss function combines the stability of L2 loss for small errors with the robustness of L1 loss for larger errors. The L1-based component is the main contributor to stabilizing the training and the prevention of *NaN* values.

- **Robust Normalization**: A specialized scaling approach that focuses on bringing the majority of data values into a standard range without excessive compression of the entire distribution. This normalization method is described in detail in section 4.5.3.

4.5.1.1. 20-Channel Samples

The TUEG pretraining dataset predominantly consists of 22-channel EEG recordings, but a small portion of samples was recorded using only 20 channels, while still following the standard 10/20 system. To ensure compatibility with the models expecting 22 channels, the following procedure was implemented:

1. Set the data in missing channels to zero.
2. Flag the missing channels in a dedicated *inherent channel mask*.
3. Explicitly exclude any such marked channel from the reconstruction loss computation.

By excluding the zero-filled missing channels from the loss calculation, the model is prevented from learning to reconstruct non-existent data.

4.5.2. Downstream Data Characteristics

Initial attempts at finetuning revealed significant challenges across the downstream datasets TUAR and TUSL. Looking at the evaluation metrics revealed consistently low precision scores, suggesting underlying class imbalance might be hampering the finetuning performance.

A detailed examination of the dataset class distributions confirmed substantial imbalances (Figure 4.17):

- **TUAB:** Nearly perfectly balanced binary classification between normal and abnormal samples.
- **TUAR:** Strong imbalance with six classes, where the majority class comprises 54.25% of samples, while the smallest class contains only 24 samples (0.05%).
- **TUSL:** Strong imbalance with the majority class representing 96.83% of all samples, while the three remaining classes each constitute approximately 1% of the dataset.

These imbalances explain the initial poor finetuning performance when training with standard cross-entropy loss, which tends to bias predictions toward the majority classes. To address this issue, Focal Loss (4.4.1.2) was employed, with class weights inversely proportional to class frequencies. This combination proved to significantly improve precision metrics of TUAR and TUSL finetuning.

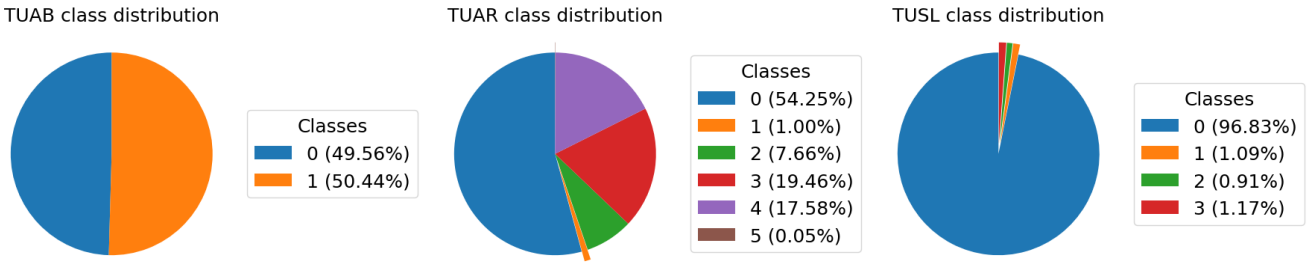


Figure 4.17.: Relative distributions of classes in TUAB, TUAR, TUSL.

4.5.3. Robust Normalization

Conventional min-max normalization would compress the entire data range into [0, 1], causing the meaningful signal variations to occupy only a tiny fraction of the normalized range due to the extreme outliers. To address this, a robust normalization approach was implemented. The approach is conceptually similar to interquartile range (IQR) normalization [48], but uses the 2nd and 98th percentiles instead of the standard 25th and 75th percentiles.

Using the 2nd and 98th percentiles across the dataset, the resulting normalization function is defined as:

$$x_{norm} = \frac{x - q_{lower}}{q_{upper} - q_{lower}}$$

where the percentile values for TUEG are defined as $(q_{lower}, q_{upper}) = (-20, 20)$. Values outside this range are allowed to exceed the [0, 1] bounds, preserving their relative magnitude.

This approach effectively brings the majority of the data (inner 96%) into a standardized range while allowing outliers to remain proportionally larger without compromising the resolution of the typical signal values.

Chapter 5

Results

5.1. Evaluation setup

The various pretraining strategies presented in previous sections were evaluated on the three models MEST, FEMBA and LaBraM. The MEST model was used for pretraining experiments with all methods, including the ones previously mentioned as no longer pursued. The FEMBA model was also pretrained using all final methods presented in the following. To gauge the benefit of doing any pretraining in the first place, the FEMBA model was also finetuned from scratch on the TUAR and TUSL datasets.

Due to computational constraints, the LaBraM model is evaluated only on a selection of the most promising methods. In addition, finetunings with the LaBraM model are also performed on the publicly available pretrained LaBraM weights from the original work [25]. It used a random masking scheme combined with a vector quantization stage and was pretrained on a different dataset, notably containing the datasets TUAR and TUSL in the pretraining data.

The pretrained models were evaluated in two ways:

- **Finetuning:** Finetuning was performed on the TUAR, TUSL and TUAB datasets, as described in section 4.4.
  Table B.5 in appendix B details the hyperparameters used for finetuning.
- **Embeddings Evaluation:** The pretrained models were used to extract embeddings from the TUAR, TUSL and TUAB datasets. These embeddings were then evaluated using linear regression classifiers to assess their quality without the influence of finetuning, and visualized using t-SNE dimension reduction.

5.1.1. Finetuning Performance

Extensive experiments were conducted comprising 18 pretraining configurations (7 strategies each for MEST and FEMBA, 4 for LaBraM) followed by finetuning runs across

three clinically relevant EEG classification tasks. For the smaller TUAR and TUSL datasets, each finetuning procedure was repeated with 5 different random seeds to ensure statistical robustness, while the larger TUAB dataset was evaluated with a single run due to computational constraints. Performance is reported using three complementary metrics established in the literature [25, 19, 7]: *Accuracy* for overall classification performance, *AUROC* for discrimination ability, and *AUPRC* for performance on imbalanced datasets.

5.1.1.1. Note on computational constraints

This work relied on two external services proving high-performance computing clusters for training. Some technical limitations have impacted the scope and completeness of the experiments. During the project, one of the two services proved unsuitable for pretraining, due to I/O speed limitations when loading the large datasets. This meant all the pretraining relied on the other service.

During the final three weeks of the project, a crucial time when a large portion of final experiments was scheduled to be performed, the remaining cluster experienced approximately two weeks of cumulative downtime due to maintenance. This necessitated a strategic prioritization of experiments:

- The LaBraM model was only evaluated on a subset of the pretraining methods, as the computational resources were limited.
- TUAB finetuning was restricted to single runs rather than the statistically preferred multiple seed.
- The sparse-attention model was not further pursued, as it would have required more experimentation and fundamental tuning to get working at all.

Despite these limitations, a total of 18 distinct pretraining configurations across three architectures and three downstream tasks could be finetuned and evaluated, providing substantial evidence for our conclusions regarding the relative effectiveness of different self-supervised learning strategies.

5.1.2. Embeddings Evaluation

To further assess the effectiveness of the pretraining strategies described above, the quality of the learned embeddings was evaluated using non-finetuning methods. This approach directly gauges how well the encoder separates EEG data without the potential compensating or altering influence of the subsequent finetuning steps.

1. **Embedding Extraction:** The pretrained encoder was used to generate embeddings for each sample in the TUAB, TUAR and TUSL datasets, omitting the decoder or classification head.

2. **Dimensionality Reduction:** For a visual inspection of class separability, a t-SNE transformation was applied to reduce each embedding to two dimensions. This 2D visualization allows for a qualitative assessment of how well samples belonging to the same class group together.

3. **Clustering Metrics:** While the 2D visualization provides insight into cluster structure, a quantitative analysis of clustering quality was also conducted. Specifically, linear regression for classification was performed directly on the encoder embeddings, using the same train-test split as finetuning uses. The classification metrics *Accuracy*, *AUROC*, and *AUPRC* were then measured.

Examining embeddings without finetuning ensures performance improvements cannot be attributed solely to a powerful classifier head or the finetuning process compensating for subpar pretraining. This methodology focuses on the raw representational ability of the encoder itself.

Visual inspection using the t-SNE plots often revealed that the embeddings showed coherent, well-defined clusters for certain classes, though some boundaries remain overlapping. This indicates that good class separation can be accomplished on the raw embeddings. Using a simple linear regression classifier ensures that the results are not influenced by an overly capable classifier head, nor is its training overly sensitive to small or imbalanced datasets.

Clustering-based evaluation of the embeddings can thus provide a clearer picture of intrinsic embedding quality across the different SSL approaches. While less relevant to the final intended use-case and not necessarily indicative of downstream performance, such analysis can be valuable when the finetuning procedure is flawed and its results might be inconsistent.

5.2. Results

5.2.1. TUSL Results

Table 5.1.: Finetuning results of MEST on TUSL dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.367 ± 0.079	0.643 ± 0.020	0.265 ± 0.009
Reconstruction - Clustered Random	0.512 ± 0.042	0.711 ± 0.023	0.275 ± 0.005
Reconstruction - Random with Lowpass	0.474 ± 0.068	0.682 ± 0.053	0.267 ± 0.007
Reconstruction - Random with JEPA	0.352 ± 0.013	0.606 ± 0.013	0.257 ± 0.003
Contrastive - Frequency	**0.570 ± 0.025**	0.673 ± 0.040	0.270 ± 0.005
Contrastive - Channel-mixed	0.412 ± 0.050	0.681 ± 0.013	0.278 ± 0.005
Contrastive - Masking	0.468 ± 0.078	**0.717 ± 0.036**	**0.279 ± 0.005**

Table 5.2.: Linear Regression results on MEST embeddings of TUSL dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.976	0.7298	0.2764
Reconstruction - Clustered Random	0.976	0.7503	**0.2978**
Reconstruction - Random with Lowpass	0.976	**0.7734**	0.2972
Reconstruction - Random with JEPA	0.976	0.7486	0.2775
Contrastive - Frequency	0.976	0.662	0.2894
Contrastive - Channel-mixed	0.976	0.7721	0.2883
Contrastive - Masking	0.976	0.7645	0.2933

Table 5.3.: Finetuning results of FEMBA on TUSL dataset

SSL Task	Accuracy	AUROC	AUPRC
No Pretraining	0.481 ± 0.053	0.666 ± 0.046	0.260 ± 0.003
Reconstruction - Random	0.485 ± 0.051	0.699 ± 0.039	0.280 ± 0.007
Reconstruction - Clustered Random	0.500 ± 0.064	0.712 ± 0.023	0.281 ± 0.019
Reconstruction - Random with Lowpass	0.519 ± 0.062	0.750 ± 0.035	0.294 ± 0.009
Reconstruction - Random with JEPA	0.463 ± 0.060	0.747 ± 0.062	0.285 ± 0.019
Contrastive - Frequency	**0.597 ± 0.072**	0.751 ± 0.020	0.291 ± 0.014
Contrastive - Channel-mixed	0.579 ± 0.050	**0.757 ± 0.019**	**0.296 ± 0.014**
Contrastive - Masking	0.441 ± 0.040	0.695 ± 0.020	0.272 ± 0.004

Table 5.4.: Linear Regression results on FEMBA embeddings of TUSL dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.976	0.727	0.267
Reconstruction - Clustered Random	0.976	0.733	0.279
Reconstruction - Random with Lowpass	0.975	0.713	0.278
Reconstruction - Random with JEPA	0.976	0.715	0.275
Contrastive - Frequency	0.976	0.730	0.274
Contrastive - Channel-mixed	0.976	0.753	0.280
Contrastive - Masking	0.976	**0.800**	**0.285**

Table 5.5.: Finetuning results of LaBraM on TUSL dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Public Weights	0.431 ± 0.015	**0.794 ± 0.013**	**0.296 ± 0.012**
Reconstruction - Random	0.451 ± 0.055	0.656 ± 0.040	0.259 ± 0.003
Reconstruction - Clustered Random	0.588 ± 0.027	0.770 ± 0.018	0.270 ± 0.002
Contrastive - Masking	**0.669 ± 0.008**	0.745 ± 0.012	0.271 ± 0.003

Table 5.6.: Linear Regression results on LaBraM embeddings of TUSL dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Public Weights	0.976	0.7126	0.271
Reconstruction - Random	0.976	0.6418	0.2823
Reconstruction - Clustered Random	0.976	**0.7605**	**0.2843**
Reconstruction - Random with JEPA	0.976	0.7329	0.2712
Contrastive - Masking	0.976	0.7116	0.2603

5.2.2. TUSL Results interpretation

The TUSL dataset presents unique challenges for EEG classification due to its small size and highly imbalanced class distribution. With the three non-base classes each only representing around 1% of the total dataset (16k samples), finetuning on this dataset is particularly difficult, as the model may struggle to learn meaningful representations for these minority classes, leading to overfitting and poor generalization. Considering this characteristic, the AUPRC is particularly relevant for results on TUSL. Besides the challenge from class imbalance, the overall small size of the dataset means there is a lot of potential to benefit from pretraining on large unlabelled data.

For the FEMBA model architecture, *Contrastive Learning with Frequency Augmentation* achieved highest accuracy, while *Contrastive Learning with Channel-Mix Augmentation* showed superior performance in both AUROC and AUPRC metrics. Similarly, for MEST, *Contrastive Learning with Frequency Augmentation* led in accuracy, while *Contrastive Learning with Mask Augmentation* showed superior performance in both AUROC and AUPRC metrics. The *public-weights* pretrained model performs best out of the LaBraM variants, with the *Clustered Masked Reconstruction* variant best out of the proposed methods.

The finetuning results on TUSL show substantial variability across the training runs with different random seeds. This variability can be attributed to the small dataset size, and limits the reliability of the finetuning results.

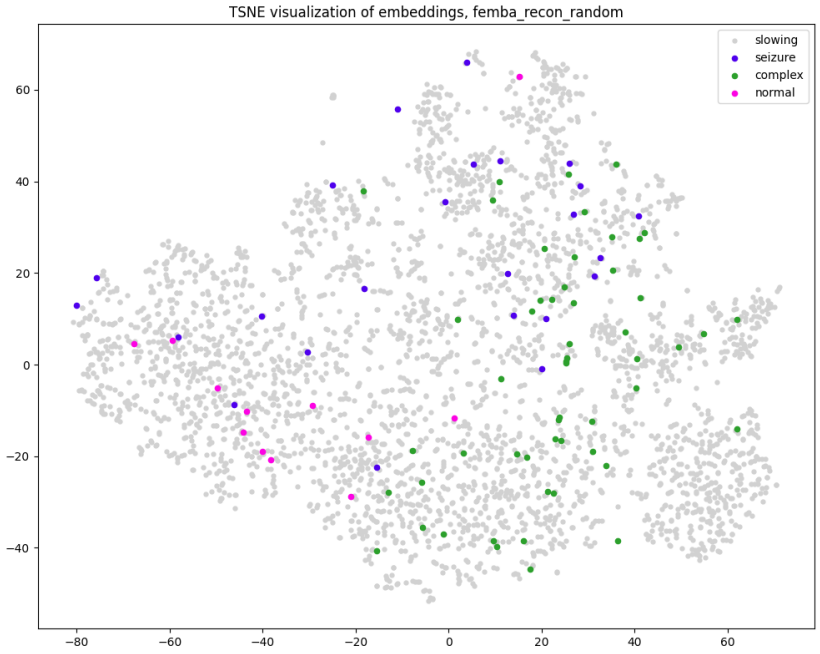
A notable observation is that the best linear regression classifiers trained directly on the embeddings from the pretrained encoder consistently outperformed their finetuned counterparts across the MEST and FEMBA architectures, for MEST on all three metrics, and for FEMBA on Accuracy and AUROC. This unexpected result suggests that the

finetuning process may have been suboptimal for this particular dataset, potentially due to the limited sample size and class imbalance. Given these circumstances, the embedding analysis provides more reliable insights into the representational quality of different pretraining strategies.

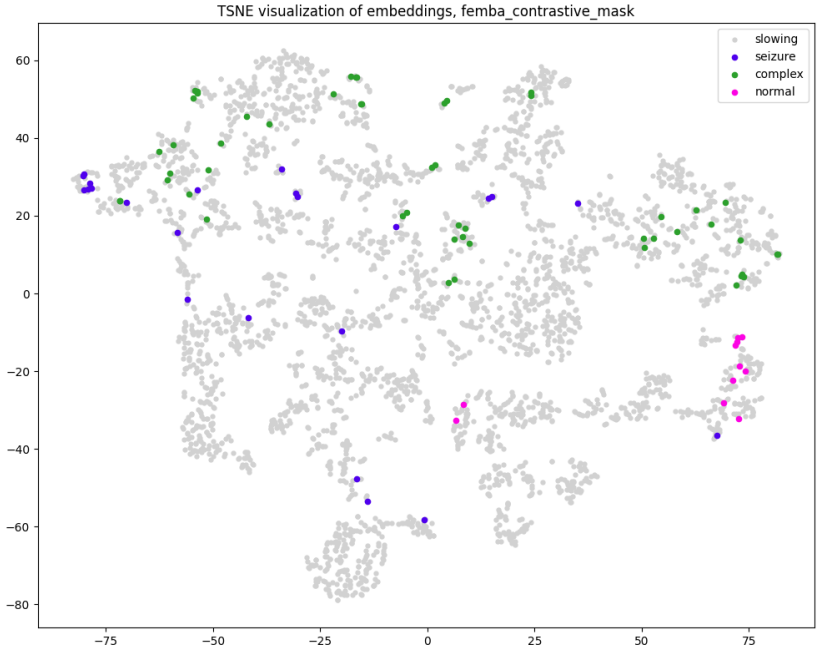
AUPRC metrics are particularly informative for the TUSL dataset given its severe class imbalance. The embeddings pretrained with *Random Masked Reconstruction* exhibit the poorest (MEST) or second-poorest (FEMBA) performance. For FEMBA, *Contrastive Learning with Mask Augmentation* delivered the best performance. Meanwhile, MEST showed strong results with *Clustered Masked Reconstruction*, followed closely by *Lowpass Target Reconstruction* and *Contrastive Learning with Mask Augmentation*. For the LaBraM model, the linear regression results reflect the finetuning performance relatively well, with finetuning achieving the higher scores. In the linear regression evaluation, LaBraM with *Clustered Masked Reconstruction* scored best on both AUROC and AUPRC.

The qualitative differences in embedding quality are visually apparent in Figure 5.1, which compares t-SNE projections of FEMBA embeddings pretrained with *Random Masked Reconstruction* and *Contrastive Learning with Mask Augmentation*. The latter demonstrates a more coherent clustering pattern, indicating better class separation.

In summary, the results suggest that contrastive learning methods are particularly effective for the TUSL dataset, outperforming reconstruction-based methods when finetuning. The findings of the embeddings analysis also shows that the baseline of *Random Masked Reconstruction* can be consistently outperformed by other pretraining methods.



(a) FEMBA TUSL embeddings with *Random Masking Reconstruction*



(b) FEMBA TUSL embeddings with *Contrastive with Mask Augmentation*

Figure 5.1.: Comparison of t-SNE embeddings of FEMBA on TUSL. *Contrastive with Mask Augmentation* shows a more coherent clustering with better class separation, while *Random Masking Reconstruction* shows a lot of overlap between the classes.

5.2.3. TUAR Results

Table 5.7.: Finetuning results of MEST on TUAR dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.566 ± 0.002	0.924 ± 0.003	0.567 ± 0.026
Reconstruction - Clustered Random	0.543 ± 0.025	0.929 ± 0.003	0.583 ± 0.011
Reconstruction - Random with Lowpass	0.567 ± 0.006	0.923 ± 0.003	0.572 ± 0.009
Reconstruction - Random with JEPA	0.561 ± 0.020	0.924 ± 0.005	0.564 ± 0.017
Contrastive - Frequency	0.577 ± 0.013	0.937 ± 0.001	0.629 ± 0.012
Contrastive - Channel-mixed	0.566 ± 0.053	**0.939 ± 0.002**	0.611 ± 0.038
Contrastive - Masking	**0.595 ± 0.007**	0.937 ± 0.001	**0.637 ± 0.019**

Table 5.8.: Linear Regression results on MEST embeddings of TUAR dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.7042	0.8973	**0.5923**
Reconstruction - Clustered Random	**0.7164**	**0.9085**	0.5861
Reconstruction - Random with Lowpass	0.6966	0.8985	0.5395
Reconstruction - Random with JEPA	0.7107	0.9007	0.5901
Contrastive - Frequency	0.6988	0.8719	0.4363
Contrastive - Channel-mixed	0.6816	0.8514	0.4792
Contrastive - Masking	0.6416	0.8221	0.4133

Table 5.9.: Finetuning results of FEMBA on TUAR dataset

SSL Task	Accuracy	AUROC	AUPRC
No Pretraining	0.470 ± 0.020	0.878 ± 0.025	0.505 ± 0.024
Reconstruction - Random	0.512 ± 0.019	0.917 ± 0.002	0.535 ± 0.014
Reconstruction - Clustered Random	0.478 ± 0.025	0.917 ± 0.004	0.547 ± 0.027
Reconstruction - Random with Lowpass	0.470 ± 0.042	0.912 ± 0.008	0.523 ± 0.026
Reconstruction - Random with JEPA	0.470 ± 0.033	0.919 ± 0.006	0.533 ± 0.013
Contrastive - Frequency	0.458 ± 0.015	**0.921 ± 0.007**	0.534 ± 0.023
Contrastive - Channel-mixed	0.493 ± 0.053	0.906 ± 0.024	**0.567 ± 0.041**
Contrastive - Masking	**0.514 ± 0.030**	0.915 ± 0.012	0.533 ± 0.014

Table 5.10.: Linear Regression results on FEMBA embeddings of TUAR dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.696	0.8701	0.5258
Reconstruction - Clustered Random	**0.727**	0.8928	0.585
Reconstruction - Random with Lowpass	0.721	**0.898**	**0.602**
Reconstruction - Random with JEPA	0.716	0.8885	0.5729
Contrastive - Frequency	0.675	0.8618	0.4683
Contrastive - Channel-mixed	0.665	0.8442	0.4089
Contrastive - Masking	0.678	0.8577	0.4462

Table 5.11.: Finetuning results of LaBraM on TUAR dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Public Weights	0.563 ± 0.026	0.918 ± 0.008	0.527 ± 0.027
Reconstruction - Random	0.474 ± 0.036	0.923 ± 0.005	0.532 ± 0.011
Reconstruction - Clustered Random	0.511 ± 0.044	0.926 ± 0.003	0.546 ± 0.028
Contrastive - Masking	**0.667 ± 0.037**	**0.940 ± 0.003**	**0.646 ± 0.041**

Table 5.12.: Linear Regression results on LaBraM embeddings of TUAR dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Public Weights	**0.7283**	**0.903**	**0.638**
Reconstruction - Random	0.6897	0.8666	0.5395
Reconstruction - Clustered Random	0.6824	0.8374	0.5295
Reconstruction - Random with JEPA	0.6963	0.8619	0.5451
Contrastive - Masking	0.6582	0.8187	0.4772

5.2.4. TUAR Results interpretation

Unlike TUSL, the TUAR dataset presents a somewhat more balanced classification task with a larger sample size (49k), though it still contains class imbalance with two seizure-related classes (*chew* and *shiv*) each representing less than 1% of the dataset.

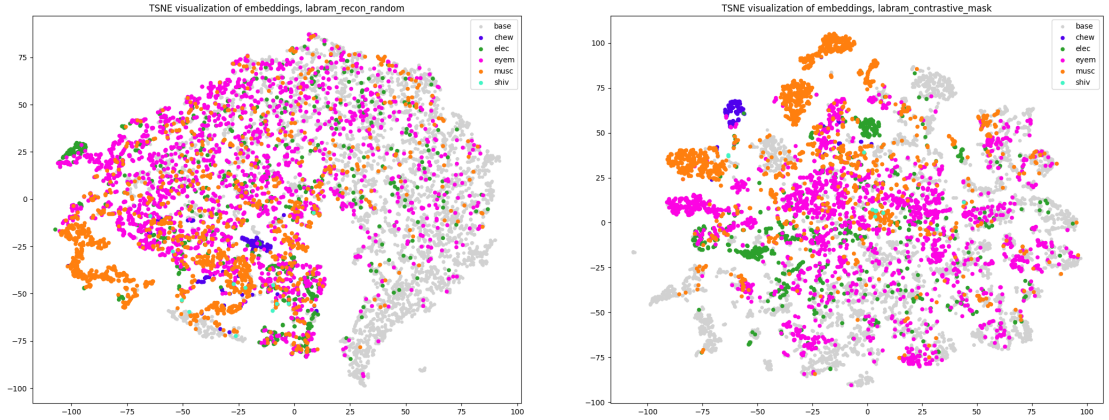
The finetuning results on TUAR demonstrate substantially lower variance across random seeds compared to TUSL, indicating more reliable performance measurements. This stability allows for more confident conclusions about the relative effectiveness of different pretraining strategies. For all three model architectures, at least one contrastive learning method consistently performed best. Specifically, FEMBA achieved its best results with *Contrastive Learning with Frequency Augmentation*, while MEST performed best with *Contrastive Learning with Channel-Mix Augmentation* and the other two *Contrastive*

methods close behind. For LaBraM, the only evaluated contrastive method *Contrastive with Masking* achieved best performance scores, while the *public weights* scored lowest. However, overall the performance differences between the pretraining methods were notably smaller than those observed in the TUSL dataset.

A stark contrast to the TUSL findings is that finetuned models significantly outperformed linear regression classifiers trained on frozen embeddings. This reversal suggests that for larger datasets like TUAR, the additional capacity provided by a trainable classification head becomes valuable, while the risk of overfitting is reduced by the increased data availability.

The t-SNE visualizations in Figure 5.2 shows a comparison of the worst and best performing pretraining methods for LaBraM on TUAR.

In summary, while the TUAR results show smaller differences between pretraining strategies compared to TUSL, contrastive learning methods still consistently outperform reconstruction approaches.



(a) LaBraM TUAR embeddings with *Random Masked Reconstruction*

(b) LaBraM TUAR embeddings with *Contrastive with Masking Augmentation*

Figure 5.2.: Comparison of t-SNE embeddings of LaBraM on TUAR. The contrastive method shows better separation between classes *elec*, *chew* and the other classes.

5.2.5. TUAB Results

Table 5.13.: Finetuning results of MEST on TUAB dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.8086	0.8730	0.8759
Reconstruction - Clustered Random	0.8053	0.8723	0.8724
Reconstruction - Random with Lowpass	**0.8114**	0.8778	**0.8834**
Reconstruction - Random with JEPA	0.8017	0.8724	0.8744
Contrastive - Frequency	0.8050	0.8718	0.8715
Contrastive - Channel-mixed	0.8045	0.8840	0.8809
Contrastive - Masking	0.8089	**0.8862**	0.8827

Table 5.14.: Linear Regression results on MEST embeddings of TUAB dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.6826	0.7436	0.7592
Reconstruction - Clustered Random	0.6978	0.7666	0.7756
Reconstruction - Random with Lowpass	**0.6944**	0.7664	**0.7783**
Reconstruction - Random with JEPA	0.6475	0.6947	0.7113
Contrastive - Frequency	0.5921	0.6152	0.6025
Contrastive - Channel-mixed	0.6878	0.7544	0.7493
Contrastive - Masking	0.7535	**0.8271**	0.8379

Table 5.15.: Finetuning results of FEMBA on TUAB dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.8171	0.8915	0.8854
Reconstruction - Clustered Random	0.8149	0.8685	0.8664
Reconstruction - Random with Lowpass	**0.8205**	**0.8930**	**0.8976**
Reconstruction - Random with JEPA	0.8178	0.8855	0.8902
Contrastive - Frequency	0.7868	0.8640	0.8620
Contrastive - Channel-mixed	0.8050	0.8679	0.8687
Contrastive - Masking	0.8077	0.8738	0.8698

Table 5.16.: Linear Regression results on FEMBA embeddings of TUAB dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Random	0.6754	0.7312	0.7279
Reconstruction - Clustered Random	0.6894	0.7504	0.7581
Reconstruction - Random with Lowpass	0.6861	0.7426	0.7468
Reconstruction - Random with JEPA	0.6277	0.6776	0.6853
Contrastive - Frequency	0.6967	0.7601	0.7410
Contrastive - Channel-mixed	**0.7230**	**0.7953**	**0.8066**
Contrastive - Masking	0.7223	0.7906	0.7912

Table 5.17.: Finetuning results of LaBraM on TUAB dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Public Weights	**0.8030**	**0.8883**	**0.8878**
Reconstruction - Random	0.7943	0.8646	0.8654
Reconstruction - Clustered Random	0.7962	0.8699	0.8702
Contrastive - Masking	0.7928	0.8699	0.8770

Table 5.18.: Linear Regression results on LaBraM embeddings of TUAB dataset

SSL Task	Accuracy	AUROC	AUPRC
Reconstruction - Public Weights	**0.7454**	**0.8227**	**0.8353**
Reconstruction - Random	0.7063	0.7697	0.7861
Reconstruction - Clustered Random	0.6973	0.7630	0.7795
Reconstruction - Random with JEPA	0.7027	0.7683	0.7775
Contrastive - Masking	0.7085	0.7880	0.7995

5.2.6. TUAB Results interpretation

The TUAB dataset is again substantially larger than TUSL and TUAR at almost 600k samples, and can be considered perfectly balanced with two classes (normal vs. abnormal EEG). For this balanced binary classification task the most relevant metric is AUROC, which best captures discriminative performance in balanced datasets. Due to computational constraints and the dataset's size, finetuning was performed with a single run rather than multiple seeds, thus no statistical significance can be reported. However, the large and balanced dataset suggests that finetuning results can be expected to be relatively stable, making individual runs more reliable indicators of performance.

For the MEST architecture, *Contrastive Learning with Mask Augmentation* achieved the highest AUROC, with *Contrastive Learning using Channel-Mix Augmentation* following

closely. The *Masked Reconstruction with Lowpass Target* method, ranking third in AUROC, achieves the highest AUPRC. The FEMBA model pretrained with *Masked Reconstruction with Lowpass Target* achieved the highest AUROC, while the *Random Masked Reconstruction* surprisingly achieved second best AUROC. The LaBraM model using the *public weights* outperformed all other methods, including the *Contrastive Learning with Mask Augementation* method, which ranked best among the novel methods, similar to the MEST results.

Considering the results of the linear regression classifiers, the *Contrastive Learning with Mask Augmentation* again ranked best for MEST and LaBraM among the novel methods, with LaBraM *public weights* again being best overall. For the FEMBA model, the *Contrastive Learning with Channel-Mix Augmentation* method achieved the best performance, with *Contrastive Learning with Mask Augmentation* following closely behind. Overall, the results on TUAB favor contrastive learning methods, especially *Contrastive Learning with Mask Augmentation*, but *Masked Reconstruction with Lowpass Target* also shows strong performance.

5.3. Discussion of pretraining methods

FEMBA without pretraining To establish a baseline for evaluating the value of pretraining in general, the FEMBA model was also finetuned directly on the TUAR and TUSL datasets without any pretraining. Across both datasets, this non-pretrained variant consistently and quite significantly underperformed compared to all pretrained counterparts, regardless of the pretraining approach employed. This considerable performance disparity underscores the critical importance of pretraining in the model's learning process, demonstrating that even basic pretraining provides substantial benefits over direct finetuning.

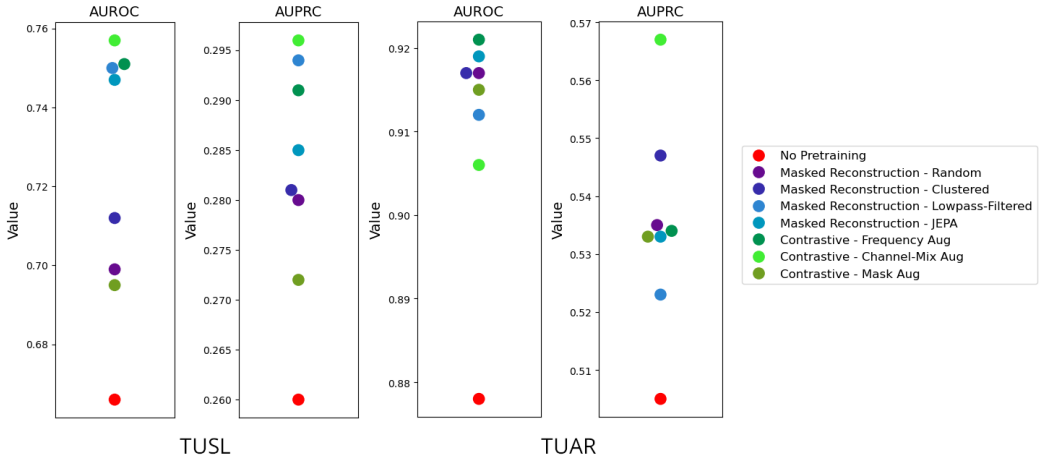
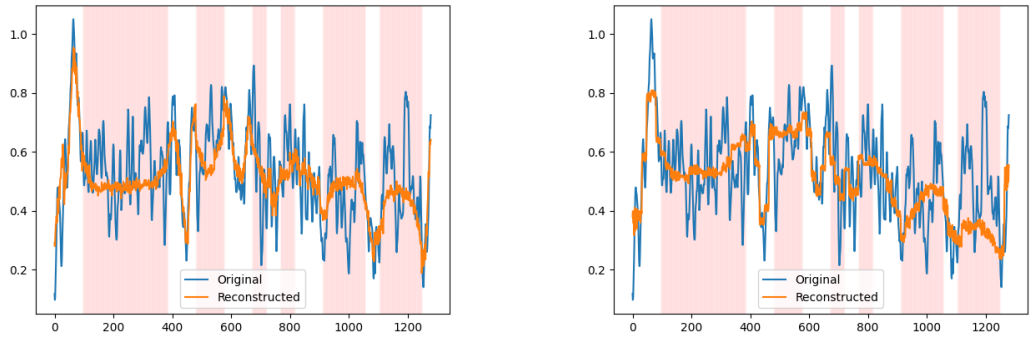


Figure 5.3.: Performance of pretraining methods on TUSL and TUAR, using FEMBA model. Pretrained methods consistently better than non-pretrained.

Random Masked Reconstruction As the current standard in most EEG foundation model pretraining, *Random Masked Reconstruction* serves as a baseline for comparison. In the experiments conducted, this approach relatively consistently ranked among the worst performing methods, and rarely was it among the upper half in terms of performance.

Clustered Random Masked Reconstruction *Clustered Random Masking* showed only marginal improvements over the standard random masking approach. While its theoretical foundation - preventing trivial signal interpolation by masking longer temporal segments - is sound, the three models employed here appear to lack sufficient capacity to extract the necessary contextual information to reconstruct in the large masked areas. This is evident when examining the reconstructions (see figure 5.4), where the model still mostly predicts an interpolation. Nevertheless, this remains a cheap method in terms of computational and implementation cost, that promises significant improvements if the model is capable of extracting the necessary contextual information, which is easily observable through inspection of the reconstructions.



(a) MEST signal reconstruction

(b) FEMBA signal reconstruction

Figure 5.4.: Signal reconstructions of MEST and FEMBA using *Clustered Random Masking*, masked regions in red. In the large masked areas the reconstruction is mostly an interpolation of the surrounding signal.

Random Masked Reconstruction with Lowpass-filtered Target The *Lowpass Target Reconstruction* often demonstrated superior performance compared to the standard random masking approach. By requiring only reconstruction of EEG-relevant frequencies, this method effectively encourages the model to learn better noise rejection. Implementation-wise, it is a simple addition to the standard random masking approach, with some caveats: First, the lowpass filtering requires either additional computation during training if performed online, or offline pre-computation at the cost of additional storage and data-loading. Second, it requires specific domain knowledge regarding the cutoff

frequency up to which relevant information can be expected, although in the case of EEG this frequency range is mostly clear.

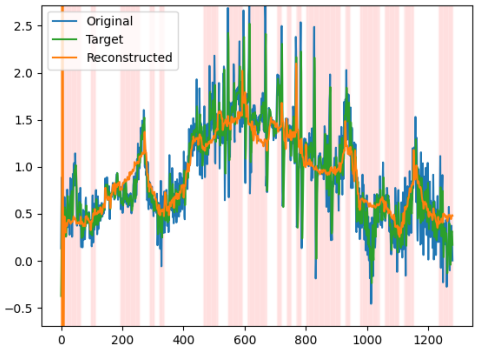


Figure 5.5.: Signal reconstruction of FEMBA using *Random Masked Reconstruction with Lowpass-filtered Target*, masked regions in red.

Random Masked Reconstruction with JEPA The *JEPA* method showed inconsistent benefits that failed to justify its additional complexity. Requiring an additional EMA student-teacher model setup, this model increases memory requirements and constrains the decoder architecture to one that can reconstruct the signal, while also being able to be interfaced at an intermediate layer for reconstruction loss at the latent dimension. A possible reason for the lack of performance has been observed during training, from logging the cosine similarity of the loss gradients of both the reconstruction and JEPA losses. The analysis revealed that the losses can at times oppose each other, demonstrating that ensemble-losses with different targets can actually oppose each other and in turn hurt performance. While this method may have some potential, it has not yet sufficiently demonstrated its benefits in the current setup of combining JEPA with conventional reconstruction.

Contrastive Learning Methods Contrastive approaches using augmentations consistently emerged as top performers across all datasets, though with some variation on which specific augmentation-type is best among the different downstream tasks. Overall, the *Channel-Mixing* and *Masking Augmentation* performed better than the *Frequency Augmentation*. Each approach offers distinct trade-off's:

- **Mask Augmentation** often provides strong performance with minimal computational overhead. Both the theoretical motivation and implementation are similar to masked reconstruction, with the benefits that contrastive methods provide.
- **Channel-Mix Augmentation** leverages spatial relationships between electrodes, but requires knowledge of the electrode layout. This methods performs especially well on the TUAR and TUSL classification tasks.

- **Frequency-based Augmentation** proved to be a relatively solid choice, though rarely the absolute best, and comes with the additional computational costs of Fourier transforms. This method depends on many different hyperparameters, further improvements can be expected with more thorough tuning.

The consistent strength of contrastive methods suggests they may better capture discriminative features in EEG signals compared to reconstruction-based approaches, particularly for classification tasks. Variations among the best suited augmentation type over different downstream tasks aligns with similar observations made by Rommel et al. [41] in their evaluation of different augmentation types. As the contrastive methods do not require any additional model components and work solely based on augmentations of the input signal, such methods may offer further potential if unified in a single SSL training strategy.

5.3.1. Discussion of Methods

From the results across different datasets using different models, a direct comparison of the different method's average performance can be drawn. To this end, a standardized analysis approach was used. Since the raw performance metrics cannot be directly averaged across different dataset and model settings, due to their different scales and performance baselines, a *z-score* standardization was applied. For each model-dataset combination, the performance metrics (Accuracy, AUROC and AUPRC) were converted to z-scores by standardizing them against all methods evaluated on that specific combination. This normalization per finetuning setting allows for meaningful aggregation across the different settings.

The resulting standardized metrics are then averaged per pretraining method, to calculate the mean z-score for the three performance indicators. This approach provides a comprehensive view of each method's relative effectiveness, accounting for varying dataset characteristics and model architectures, and are presented in Figure 5.6.

Several key findings emerge from this comparison. First, all proposed methods on average demonstrated superior performance in terms of *AUROC* and *AUPRC* compared to the baseline of conventional *Random Masked Reconstruction*.

Among the reconstruction-based methods, the *Lowpass Target Reconstruction* generally achieved the best results, suggesting that focusing the reconstruction on the relevant frequency bands improves representation learning. This method was least performant on the TUAR artifact classification task. A reason for this shortcoming may be found in the assumption underlying the lowpass-filtering, which only expects clinically relevant information in the frequencies up to 50Hz. The TUAR classification task however, focusses not on the underlying EEG signals, but on the artifacts that are corrupting them. These artifacts may contain information also in the higher frequencies, thus ignoring them might be suboptimal for this specific task.

The JEPA approach, despite its theoretical appeal, did not demonstrate competitive performance relative to its implementation complexity. The implementation used in this work differs from the original JEPA paradigm, which was used as a stand-alone

training method, by combining the JEPA loss with a conventional reconstruction loss. Such ensemble losses that are pursuing different objectives simultaneously may lead to a suboptimal model that accomplishes neither task very well. An indication that the two losses were not pursuing similar objectives was seen during observation of the loss-gradients during training. Analysing the cosine-similarity of the gradients of the two losses revealed that they were at times opposing each other, thus potentially preventing the training process from finding an optimal loss minimum for both objectives.

Most notably, the contrastive methods with *Channel-Mixing Augmentation* and *Mask Augmentation* emerged as the strongest performers overall. Both methods yield pre-trained encoders producing embeddings that qualitatively demonstrate good class separation. Contrastive methods not only encourage good feature extraction but mainly emphasize a well-separated latent space, which is conducive to downstream classification. The *Mask Augmentation* method specifically builds on the same assumptions as conventional *Masked Reconstruction*, which has proven a solid pretraining strategy over a variety of domains, without requiring the explicit reconstruction of noise in the input signal, thus potentially leading to a model that is more noise robust as well as mapping to a better separated latent space. These observations confirm the patterns identified observed in the individual dataset analyses and suggest that contrastive approaches may be particularly well-suited for capturing the discriminative features needed for EEG classification tasks.

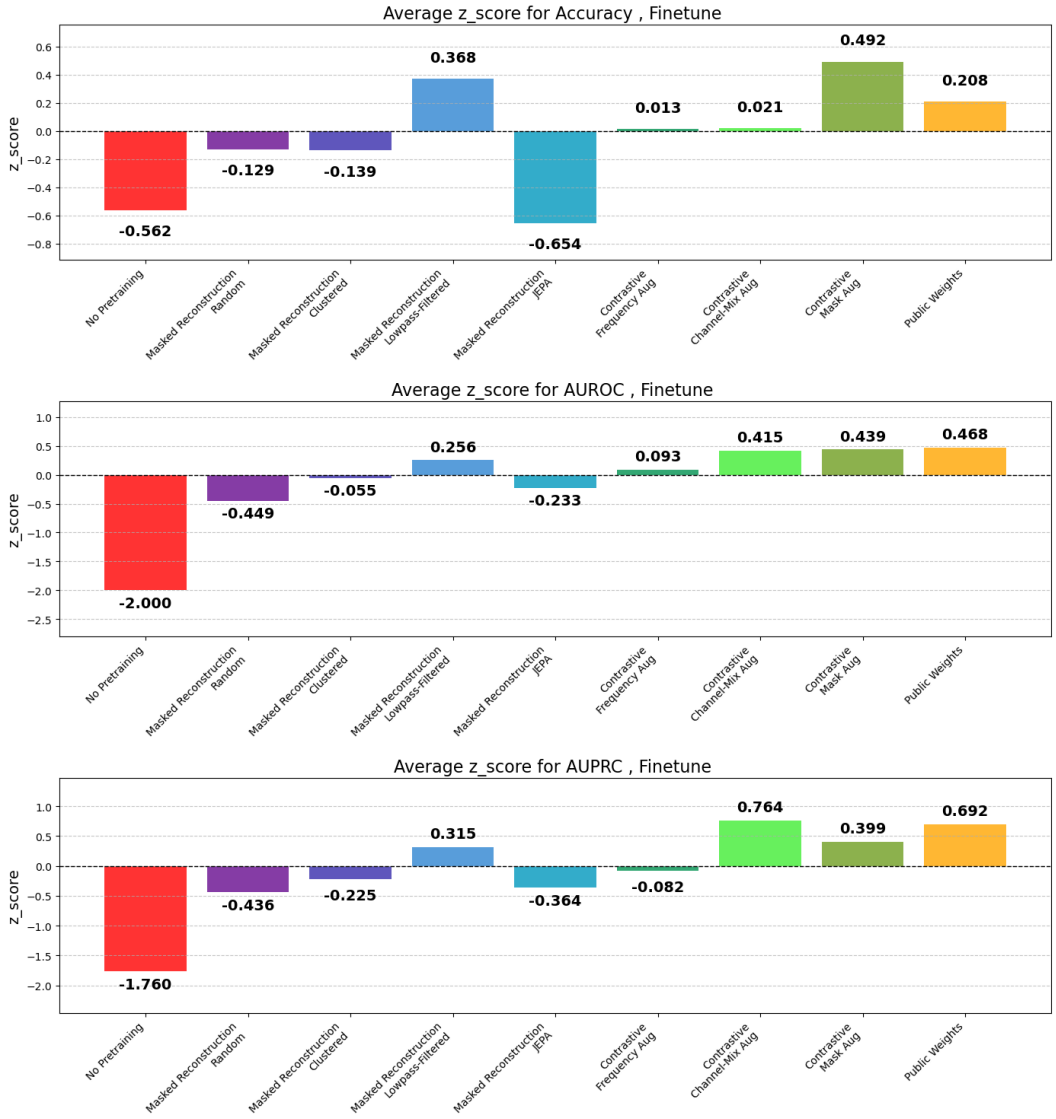


Figure 5.6.: Averaged z-score-standardized metrics of all method's finetuning performance compared. Standardized across all methods per model-dataset combination.

5.4. Current limitations

While this work provides insights into self-supervised learning strategies for EEG representation learning, several limitations must be acknowledged:

Experimental Robustness The TUSL finetuning results demonstrate substantial variability across the random seed finetuning experiments with respect to the performance

metrics. This inconsistency limits the confidence with which conclusions can be drawn on downstream performance of the methods on this particular dataset.

For the TUAB dataset, computational constraints necessitated single finetuning runs. Although the large, balanced nature of TUAB mitigates some concern about run-to-run variability, the absence of multiple runs prevents proper assessment of statistical significance when comparing methods on this downstream task.

All pretraining runs were also conducted with single seeds, rather than multiple initializations. In this case however, this limitation is mitigated by several factors: the large pretraining data volume (15M+ signal samples), substantial effective batch sizes (2400 using gradient accumulation), and the consistent convergence patters observed across many runs. These factors tend to reduce the impact of random initialization on the final model performance.

LaBraM Method Coverage The LaBraM model architecture was evaluated on only a subset of the pretraining methods. Specifically, methods *Random Reconstruction with Lowpass-filtered Target*, *Contrastive with Frequency Augmentation* and *Contrastive with Channel-Mix Augmentation* were not evaluated. This limited coverage may affect the generalizability of the findings, as these methods have shown strong performance in other architectures.

Additionally, preliminary comparisons suggest that the publicly available weights of the LaBraM model can outperform the pretrained models from this work. Specifically, the *Public Weights* model performs significantly better than the *Random Masking* model, with which it shares the masked reconstruction pretraining strategy. The results further show that while the *Public Weights* are not generally the best after finetuning, they do clearly score highest using linear regression classification, indicating superior linear separability. Two different factors may contribute to this divergence in performance. The model-specific pretraining using vector-quantization in the original LaBraM work may be contributing to the higher downstream performance, independent of the masked reconstruction pretraining strategy. Furthermore, these publicly available LaBraM weights were also pretrained on a different, diverse dataset, which included the TUAR and TUSL datasets used here for the evaluation, thus possibly distorting the comparison.

Depth of Hyperparameter Explorations While reasonable hyperparameter configurations were selected for all models, they were mostly based on values established in literature and experimental observations of the reconstructions and limited finetuning performances. Due to the high number of methods evaluated, as well as multiple model architectures and downstream tasks, exhaustive hyperparameter searches were not feasible. The choices made for such values are thus not necessarily optimal, and the relative ranking of methods might shift with more thorough optimization.

These limitations suggest several avenues for future work, including more robust statistical evaluation, expanded architectural comparisons, and deeper hyperparameter optimization, particularly for the most promising contrastive learning approaches identified in this study.

Chapter 6

Conclusion and Future Work

This thesis has investigated self-supervised learning approaches for EEG representation learning, evaluating various pretraining strategies across multiple model architectures and downstream clinical tasks. The systematic comparison provides valuable insights into which methods produce high-quality representations for EEG data analysis.

Implementation of several EEG specific SSL strategies and evaluation on three relevant downstream tasks revealed that the performance of conventional *Random Masked Reconstruction* can be surpassed. Specifically, contrastive learning methods yielded superior performance on the downstream classification task, presumably by producing a well-separated latent space with better noise rejection than reconstructive methods. Absolute best performances were achieved with contrastive methods *Channel-Mix Augmentation* and *Masking Augmentation*, although with varying rankings across tasks and models. Reconstruction-based approaches demonstrated more consistent performance, with *Lowpass Target Reconstruction* showing particular promise. The main challenges of this work included designing appropriate domain-specific masking-strategies and augmentations, as well as addressing EEG's unique temporal and spatial characteristics. The methods have been evaluated on SoA model architectures FEMBA and LaBraM, and demonstrated that further improvements over their reported performance are possible using an optimized pretraining strategy (FEMBA with *Contrastive Channel-Mixing*: +4.9% AUROC on TUSL), though vector-quantization (as used in public LaBraM weights) provides additional benefits. Finetuning FEMBA from scratch highlighted the critical importance of pretraining in general for maximizing performance. Dataset-specific challenges, including outliers in pretraining data and class imbalances in downstream tasks, were effectively addressed through robust normalization strategies and the implementation of focal loss for finetuning, as well as a secondary evaluation using linear regression on frozen encoder embeddings.

Future work should explore several promising directions: combinations of the most effective methods identified (*Lowpass Target Reconstruction* lends itself well); contrastive learning with multiple simultaneous augmentations; incorporation of vector quantization similar to that used in the original LaBraM work; and more extensive hyperparam-

eter optimization for the most promising approaches. Utilizing sparse attention with an optimized spatio-temporal attention mask was briefly investigated, following up on this paradigm might yield a more capable yet memory-efficient model architecture for EEG processing. Additionally, the limitations acknowledged regarding experimental robustness and coverage suggest the need for broader evaluations with multiple initialization seeds and expanded architectural comparisons to further validate and extend these findings.

Appendix A

Task Description

Master Project at the Department of
Information Technology and Electrical Engineering

Autumn Semester 2024

Nicolas Lehmann

Deep Learning Based EEG Signal Understanding Using Foundational Models

September 26, 2024

Advisors: Thorir Mar Ingolfsson, thoriri@iis.ee.ethz.ch
Dr. Yawei Li, yawei.li@vision.ee.ethz.ch

Supervisor: Prof. Dr. Luca Benini, lbenini@iis.ee.ethz.ch

Handout: September 17, 2024
Due: March 18, 2025

Electroencephalography (EEG) is a non-invasive method used to record electrical activity in the brain, playing a critical role in both neurological research and clinical diagnostics [1]. By offering a window into the brain's activity, EEG helps diagnose and treat various neurological disorders. However, the analysis of EEG signals presents substantial challenges due to the complexity of these signals and the fine distinctions required between normal and abnormal brain activity [2].

In recent years, AI and deep learning have opened up new possibilities for analyzing EEG data. Foundation models—large, pre-trained neural networks that can be fine-tuned for specific tasks—have reshaped fields such as natural language processing and computer vision [3]. Their potential to improve EEG analysis is promising but still in its early stages.

Progress has been made in this area with models such as LaBraM, which segments EEG signals into channel patches to enable cross-dataset learning and fine-tuning [4]. Additionally, Brant-X offers a unified physiological signal alignment framework by leveraging EEG alongside other physiological signals, improving performance in EEG classification tasks like sleep stage and emotion recognition [5]. Another model, EEGFormer, takes a self-supervised learning approach with vector quantization to pre-train on large-scale EEG datasets, producing state-of-the-art results in tasks like seizure detection [6].

Recent advancements in self-supervised learning have opened up significant possibilities for pre-training foundation models in EEG analysis. One such model, FoME (Foundation Model for EEG), introduces adaptive temporal-lateral attention scaling and is pre-trained on a large-scale EEG dataset using a combination of time-frequency fusion and attention mechanisms, achieving state-of-the-art results in multiple downstream tasks [7].

The use of masked autoencoders, as seen in recent studies, has proven to be an effective self-supervised approach for EEG representation learning. This involves masking portions of the input signal and training the model to reconstruct the missing parts, enabling it to learn richer representations of EEG data [8]. Techniques like MAEEG (Masked Auto-Encoder for EEG) have further demonstrated how masked autoencoding can enhance representation learning for EEG data, showing potential advantages over fully supervised learning [8].

Additionally, contrastive learning has emerged as another powerful self-supervised method. For instance, BENDR (Bidirectional Encoder Representations from Transformers for EEG) utilizes contrastive learning to train on massive EEG datasets, offering effective feature extraction and transfer learning capabilities for downstream tasks [8].

These self-supervised methods offer different ways of training foundation models, but the trade-offs between them remain an active area of research, particularly when it comes to deciding between approaches like masking raw signals versus spectrograms, or learning from contrastive tasks. This project will explore these aspects in depth, aiming to identify the most effective pre-training strategy for developing robust foundation models for EEG analysis.

1 Project Description

The objective of this project is to develop and pre-train a foundational model for EEG analysis, with a particular emphasis on exploring self-supervised learning tasks. This involves investigating and comparing various self-supervised training techniques, such as masked autoencoding, contrastive learning, and other potential methods, to identify the most effective strategy for extracting robust and transferable EEG representations. The main tasks of this project are outlined as follows:

- **Task I: Literature Review and Methodology Exploration** Conduct a comprehensive review of self-supervised learning techniques used in foundation models, with a focus on their application to EEG analysis. Study existing models like FoME, MAEEG, and BENDR, paying close attention to how they implement self-supervised tasks such as masking, contrastive learning, and temporal-spatial representation learning. Identify the strengths and weaknesses of different approaches in handling EEG data.
- **Task II: Data Preparation and Preprocessing**
  Select and preprocess a large-scale EEG dataset for pre-training. This involves handling variations in EEG signals, filtering noise, segmenting the signals into appropriate windows, and possibly converting them into different formats (e.g., raw signals, spectrograms, or other time-frequency representations) to assess which format is best suited for the self-supervised tasks.
- **Task III: Experimenting with Self-Supervised Pre-Training Tasks**
  Implement various self-supervised training tasks, starting with masked autoencoders. Experiment with masking different portions of the EEG signals (e.g., raw signal segments, frequency bands) and explore the trade-offs between masking future versus past signals. Investigate how these choices impact the model's ability to learn meaningful EEG representations. Additionally, test contrastive learning tasks, comparing positive and negative pairs of EEG segments, to evaluate their effectiveness in representation learning.
- **Task IV: Investigating Multi-Modal Data Integration** Explore incorporating additional physiological data modalities alongside EEG to enhance the foundation model's representation learning capabilities. Identify and procure datasets that include complementary signals such as Electrocardiography (ECG), Electromyography (EMG), Photoplethysmogram (PPG), or other relevant modalities. Preprocess these additional data sources to ensure synchronization and alignment with the EEG signals, addressing challenges like differing sampling rates and signal characteristics. Experiment with various multi-modal fusion strategies at different levels—such as early fusion at the input stage, mid-level fusion during feature extraction, or late fusion at the decision-making stage. Evaluate the impact of integrating multiple modalities on the performance of self-supervised learning tasks, assessing improvements in representation quality and downstream task performance.

Analyze the trade-offs involved, including computational complexity and the potential for overfitting, to determine the most effective approach for multi-modal integration in EEG foundation models.

- **Task V: Developing a Self-Supervised Foundation Model**
  Design and implement a foundation model that integrates the most promising self-supervised learning techniques identified in Task III. This model should be capable of learning from the pre-training data and be adaptable to various downstream tasks. Incorporate elements like adaptive attention mechanisms or time-frequency fusion if relevant, inspired by models such as FoME or BENDR.
- **Task VI: Pre-Training the Foundation Model**
  Train the foundation model on the pre-processed EEG dataset using the selected self-supervised tasks. Monitor the training progress, evaluate the quality of the learned representations, and make adjustments to the training procedure or architecture as necessary. Experiment with combining multiple self-supervised tasks to enhance the model's ability to capture the complexity of EEG signals.
- **Task VII: Fine-Tuning and Evaluation on Downstream Tasks**
  Fine-tune the pre-trained foundation model on several downstream EEG tasks such as emotion recognition, seizure detection, or sleep stage classification. Compare the performance against existing models, analyzing how well the self-supervised pre-training contributes to task-specific performance.
- **Task VIII: Comparative Analysis and Reporting**
  Compare the results of the pre-trained foundation model with other state-of-the-art EEG models. Assess the effectiveness of different self-supervised learning strategies and their impact on the foundation model's performance across multiple tasks. Summarize the findings, including the strengths and limitations of each approach, and suggest potential improvements.
- **Task IX: Final Presentation and Documentation**
  Prepare a comprehensive report detailing the methodology, experiments, and results. Present the findings, focusing on how self-supervised learning can be leveraged to create a strong foundation model for EEG analysis. Highlight key insights, challenges, and future directions for self-supervised pre-training in EEG research.

2 Project Realization

2.1 Meetings

Weekly meetings and reports must be held. The exact time and location of these meetings will be determined within the first week of the project in order to fit the student's and the assistant's schedule. These meetings will be used to evaluate the status and progress

of the project. Besides these regular meetings, additional meetings can be organized to address urgent issues as well.

2.2 Report

Documentation is an important and often overlooked aspect of engineering. One final report has to be completed within this project. The common language of engineering is de facto English. Therefore, the final report of the work is preferred to be written in English. Any form of word processing software is allowed for writing the reports, nevertheless, the use of LaTeX with Tgif[1] or any other vector drawing software (for block diagrams) is strongly encouraged by the IIS staff.

Final Report The final report has to be presented at the end of the project and a digital copy need to be handed in. Note that this task description is part of your report and has to be attached to your final report.

2.3 Presentation

There will be a presentation (15 min for the semester thesis, and 20 min for the MS thesis presentation followed by 5 min Q&A) at the end of this project in order to present your results to a wider audience. The exact date will be determined towards the end of the work.

References

[1] S. Weisdorf, S. W. Gangstad, J. Duun-Henriksen, K. S. S. Mosholt, and T. W. Kjær, "High similarity between eeg from subcutaneous and proximate scalp electrodes in patients with temporal lobe epilepsy," *Journal of Neurophysiology*, vol. 120, no. 3, pp. 1451–1460, 2018.

[2] T. M. Ingolfsson, S. Benatti, X. Wang, A. Bernini, P. Ducouret, P. Ryvlin, S. Beniczky, L. Benini, and A. Cossettini, "Minimizing artifact-induced false-alarms for seizure detection in wearable eeg devices with gradient-boosted tree classifiers," *Scientific Reports*, vol. 14, no. 1, p. 2980, 2024.

[3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[1] See: `http://bourbon.usc.edu:8001/tgif/index.html` and `http://www.dz.ee.ethz.ch/en/information/how-to/drawing-schematics.html`.

[4] W.-B. Jiang, L.-M. Zhao, and B.-L. Lu, "Large brain model for learning generic representations with tremendous eeg data in bci," *arXiv preprint arXiv:2405.18765*, 2024.

[5] D. Zhang, Z. Yuan, J. Chen, K. Chen, and Y. Yang, "Brant-x: A unified physiological signal alignment framework," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 4155–4166.

[6] Y. Chen, K. Ren, K. Song, Y. Wang, Y. Wang, D. Li, and L. Qiu, "Eegformer: Towards transferable and interpretable large-scale eeg foundation model," *arXiv preprint arXiv:2401.10278*, 2024.

[7] E. Shi, K. Zhao, Q. Yuan, J. Wang, H. Hu, S. Yu, and S. Zhang, "Fome: A foundation model for eeg using adaptive temporal-lateral attention scaling," *arXiv preprint arXiv:2409.12454*, 2024.

[8] Y. Zhou and S. Liu, "Enhancing representation learning of eeg data with masked autoencoders," in *International Conference on Human-Computer Interaction*. Springer, 2024, pp. 88–100.

Zurich, September 26, 2024 Prof. Dr. Luca Benini

Appendix B

Training Hyperparameters

B.1. Models

Hyperparameter	MEST	FEMBA	LaBraM
Total Parameters	3.3M	7.8M	5.8M
Patch Size	32	16	200
Embedding Dimension	160	35	200
Num. Attn Heads	10	-	10
Num. Transf. Layers	10	-	12
Num. Mamba Blocks	-	2	-

Table B.1.: Model Architecture Hyperparameters for different models.

B.2. Pretraining

Hyperparameter	Value
Reconstruction	
$\lambda_{reconstruction}$	1.0
$\lambda_{unmasked\_reconstruction}$	0.1
JEPA	
λ_{JEPA}	0.1
EMA decay	0.9
Contrastive	
$\lambda_{Contrastive}$	1.0
Max freq. (FT-Surrogate)	100Hz
Max freq. (Freq. Shift)	100Hz
Max freq. shift	3Hz
Noise std	0.1
Random scaling factor (Channel-Mix)	$[0.5, 1.5]$
Max random bias mag.	1.0

Table B.2.: SSL method specific hyperparameters

Hyperparameter	MEST	FEMBA	LaBraM
Optimizer		AdamW	
Batch Size	300	600	600
Gradient Accumulation	8	4	4
Epochs		150	
Grad. Clip		1.0	
LR Scheduler		CosineAnnealingLR	
LR max		5e-4	
LR min		1e-6	
Warmup Epochs		5	
Precision		bf16-mixed	

Table B.3.: Pretraining Hyperparameters for different models.

Channel	Neighbors		
Fp1-F7	F7-T3	FP1-F3	F3-C3
F7-T3	FP1-F7	A1-T3	T3-C3
T3-T5	A1-T3	T3-C3	T5-O1
T5-O1	T3-T5	C3-P3	P3-O1
FP2-F8	FP2-F4	F4-C4	F8-T4
F8-T4	T5-O1	C4-T4	T4-A2
T4-T6	C4-T4	T4-A2	T6-O2
T6-O2	T4-T6	C4-P4	P4-O2
A1-T3	FP1-F7	F7-T3	T3-C3
T3-C3	A1-T3	C3-CZ	F3-C3
C3-CZ	F3-C3	C3-P3	CZ-C4
CZ-C4	C3-CZ	F4-C4	C4-P4
C4-T4	CZ-C4	F4-C4	T4-A2
T4-A2	C4-T4	F8-T4	T4-T6
FP1-F3	FP1-F7	F3-C3	FP2-F4
F3-C3	FP1-F3	T3-C3	C3-CZ
C3-P3	T3-C3	C3-CZ	P3-O1
P3-O1	T5-O1	C3-P3	P4-O2
FP2-F4	FP1-F3	FP2-F8	F4-C4
F4-C4	FP2-F4	CZ-C4	C4-T4
C4-P4	CZ-C4	C4-T4	P4-O2
P4-O2	T6-O2	C4-P4	P3-O1

Table B.4.: Channel-Mix augmentation neighbor definitions, 10/20 system.

B.3. Finetuning

Hyperparameter	**MEST**	**FEMBA**	**LaBraM**
Optimizer		AdamW	
Batch Size		256	
Epochs		50	
Grad. Clip		1.0	
LR	5e-5	5e-4	5e-5
LwLRD		0.7	
Weight Decay		0.05	
Dropout	-	0.0	-
Droppath	0.1	-	0.1
Noise augmentation prob		0.2	
Noise augmentation mag		0.1	
Label smoothing		0.1	
	TUAB	**TUAR**	**TUSL**
Loss	Cross-Entropy	Focal	Focal
Class Weights	$[1.0, 1.0]$	$[0.46, 0.99, 0.92, 0.81, 0.82, 1.0]$	$[0.01, 0.9, 1.1, 0.85]$

Table B.5.: Finetuning Hyperparameters for different Models and Datasets.

List of Figures

List of Tables

Bibliography

[1] V. Shah, E. Von Weltin, S. Lopez, J. R. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, and J. Picone, "The temple university hospital seizure detection corpus," *Frontiers in neuroinformatics*, vol. 12, p. 83, 2018.

[2] 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

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[6] J. Wang, S. Zhao, Z. Luo, Y. Zhou, H. Jiang, S. Li, T. Li, and G. Pan, "Cbramod: A criss-cross brain foundation model for eeg decoding," *arXiv preprint arXiv:2412.07236*, 2024.

[7] A. Tegon, T. M. Ingolfsson, X. Wang, L. Benini, and Y. Li, "Femba: Efficient and scalable eeg analysis with a bidirectional mamba foundation model," 2025. [Online]. Available: https://arxiv.org/abs/2502.06438

[8] H. H. Jasper, "Ten-twenty electrode system of the international federation," *Electroencephalogr Clin Neurophysiol*, vol. 10, pp. 371–375, 1958.

[9] D. P. Subha, P. K. Joseph, R. Acharya U, and C. M. Lim, "Eeg signal analysis: a survey," *Journal of medical systems*, vol. 34, pp. 195–212, 2010.

[10] K. A. I. Aboalayon, M. Faezipour, W. S. Almuhammadi, and S. Moslehpour, "Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation," *Entropy*, vol. 18, no. 9, p. 272, 2016.

[11] D. Wulsin, J. Blanco, R. Mani, and B. Litt, "Semi-supervised anomaly detection for eeg waveforms using deep belief nets," in *2010 Ninth international conference on machine learning and applications*. IEEE, 2010, pp. 436–441.

[12] T. M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, and L. Benini, "Eeg-tcnet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 2958–2965.

[13] J. Kaiser, J. Perelmouter, I. H. Iversen, N. Neumann, N. Ghanayim, T. Hinterberger, A. Kübler, B. Kotchoubey, and N. Birbaumer, "Self-initiation of eeg-based communication in paralyzed patients," *Clinical Neurophysiology*, vol. 112, no. 3, pp. 551–554, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245701004709

[14] R. Bommasani, D. A. Hudson, E. Adeli, R. B. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. E. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, and et al., "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021. [Online]. Available: https://arxiv.org/abs/2108.07258

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[16] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2022. [Online]. Available: https://arxiv.org/abs/2104.05704

[17] 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

[18] 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

[19] A. Dimofte, G. A. Bucagu, T. M. Ingolfsson, X. Wang, A. Cossettini, L. Benini, and Y. Li, "Cerebro: Compact encoder for representations of brain oscillations using efficient alternating attention," 2025. [Online]. Available: https://arxiv.org/abs/2501.10885

[20] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2024. [Online]. Available: https://arxiv.org/abs/2312.00752

[21] N. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, "A dataset of neonatal eeg recordings with seizure annotations," *Scientific Data*, vol. 6, p. 190039, 03 2019.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," 2022, pp. 16 000–16 009. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper

[23] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A Simple Framework for Masked Image Modeling," Apr. 2022, arXiv:2111.09886. [Online]. Available: http://arxiv.org/abs/2111.09886

[24] C. Wang, V. Subramaniam, A. U. Yaari, G. Kreiman, B. Katz, I. Cases, and A. Barbu, "Brainbert: Self-supervised representation learning for intracranial recordings," *arXiv preprint arXiv:2302.14367*, 2023.

[25] 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

[26] Y. Chen, K. Ren, K. Song, Y. Wang, Y. Wang, D. Li, and L. Qiu, "Eegformer: Towards transferable and interpretable large-scale eeg foundation model," 2024. [Online]. Available: https://arxiv.org/abs/2401.10278

[27] H.-Y. S. Chien, H. Goh, C. M. Sandino, and J. Y. Cheng, "Maeeg: Masked auto-encoder for eeg representation learning," *arXiv preprint arXiv:2211.02625*, 2022.

[28] Y. Bao, S. Sivanandan, and T. Karaletsos, "Channel vision transformers: An image is worth 1 x 16 x 16 words," 2024. [Online]. Available: https://arxiv.org/abs/2309.16108

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," 2021. [Online]. Available: https://arxiv.org/abs/2111.06377

[30] Z. Huang, X. Jin, C. Lu, Q. Hou, M.-M. Cheng, D. Fu, X. Shen, and J. Feng, "Contrastive masked autoencoders are stronger vision learners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2506–2517, 2023.

[31] P. Guetschel, T. Moreau, and M. Tangermann, "S-JEPA: Towards seamless cross-dataset transfer through dynamic spatial attention," Oct. 2024.

[32] S. Kim, "Learning General Representation of 12-Lead Electrocardiogram with a Joint-Embedding Predictive architecture," Oct. 2024.

[33] R. Girshick, "Fast r-cnn," 2015. [Online]. Available: https://arxiv.org/abs/1504.08083

[34] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "Simmim: A simple framework for masked image modeling," 2022. [Online]. Available: https://arxiv.org/abs/2111.09886

[35] P. Tangkraingkij, "Significant frequency range of brain wave signals for authentication," in *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2015*, R. Lee, Ed. Cham: Springer International Publishing, 2016, pp. 103–113.

[36] G. Buzsáki and A. Draguhn, "Neuronal oscillations in cortical networks," *Science*, vol. 304, no. 5679, pp. 1926–1929, 2004. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1099745

[37] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas, "Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture," Apr. 2023.

[38] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019. [Online]. Available: https://arxiv.org/abs/1807.03748

[39] W. Li, H. Li, X. Sun, H. Kang, S. An, G. Wang, and Z. Gao, "Self-supervised contrastive learning for EEG-based cross-subject motor imagery recognition," *Journal of Neural Engineering*, vol. 21, no. 2, p. 026038, Apr. 2024.

[40] L. Yang and S. Hong, "Unsupervised Time-Series Representation Learning with Iterative Bilinear Temporal-Spectral Fusion," May 2022.

[41] C. Rommel, J. Paillard, T. Moreau, and A. Gramfort, "Data augmentation for learning predictive models on EEG: A systematic comparison," *Journal of Neural Engineering*, vol. 19, no. 6, p. 066020, Dec. 2022.

[42] J. T. C. Schwabedal, J. C. Snyder, A. Cakmak, S. Nemati, and G. D. Clifford, "Addressing class imbalance in classification problems of noisy signals by using fourier transform surrogates," 2019. [Online]. Available: https://arxiv.org/abs/1806.08675

[43] C. Rommel, T. Moreau, J. Paillard, and A. Gramfort, "Cadda: Class-wise automatic differentiable data augmentation for eeg signals," 2022. [Online]. Available: https://arxiv.org/abs/2106.13695

[44] Tomaton124, "21 electrodes of international 10-20 system for eeg, modified," 2025, accessed: 2025-04-11. [Online]. Available: https://commons.wikimedia.org/wiki/File:21_electrodes_of_International_10-20_system_for_EEG.svg

[45] O. Yonay, T. Hammond, and T. Yang, "Myna: Masking-based contrastive learning of musical representations," 2025. [Online]. Available: https://arxiv.org/abs/2502.12511

[46] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," 2019. [Online]. Available: https://arxiv.org/abs/1803.05407

[47] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," 2018. [Online]. Available: https://arxiv.org/abs/1708.02002

[48] F. Dekking, C. Kraaikamp, H. Lopuhaä, and L. Meester, *A Modern Introduction to Probability and Statistics, Understanding Why and How*, 01 2005.