
EEG2VEC: A unified representation for EEG signals

Hazim Bukhari

Language Technology Institute
Carnegie Mellon University
hbukhari@andrew.cmu.edu

Pranav Kini

Department of Biomedical Engineering
Carnegie Mellon University
pranavki@andrew.cmu.edu

Mayur S Gowda

Electrical and Computer Engineering
Carnegie Mellon University
mayursri@andrew.cmu.edu

Abstract

Brain-computer interfaces (**BCIs**) enable users to control computers via brain signals recorded with electroencephalography (**EEG**). EEG signals typically suffer from a low signal-to-noise ratio (**SNR**), making robust denoising and feature extraction essential. Traditional methods rely on handcrafted, task specific features that often fail to generalize across paradigms. More recently, deep learning models trained for individual tasks have demonstrated improved feature extraction, but still require costly retraining to adapt to new applications.

In this paper, we introduce an end-to-end masked prediction, self-supervised foundation model for EEG that learns unified representations transferable to any downstream task without task specific fine tuning. We pretrain our model on over 8,000 hours of unlabeled EEG recordings and adapt the HuBERT architecture for multichannel EEG feature extraction. We evaluate these representations on the P300 detection task, benchmarking them against three established baselines. Although our HuBERT based model did not exceed baseline performance, we analyze the underlying causes and propose remedies for future work (see Section 6.2).

Our main contributions are:

- Pretraining on a large scale EEG corpus of over 8,000 hours.
- Adapting HuBERT to handle multichannel, low **SNR** EEG data.
- Benchmarking transferability to the P300 detection task and diagnosing performance limitations.

1 Introduction

Non-intrusive electroencephalograms (**EEG**) are electrical signals produced by neurons in the brain, measured by placing a cap with sensors on the subject’s scalp. These signals provide valuable insights into the activities inside the human brain and have extensive applications in brain-computer interfaces (**BCIs**) and healthcare domains. Although EEG signals are easy to collect, they are also sensitive to noise from the environment and eye movements. The introduced noise degrades the usefulness of the EEG signals, which in turn affects performance on downstream tasks. Recent research has leveraged self-supervised learning to learn meaningful representations for downstream tasks, including seizure detection [14] and abnormality detection [5] [17]. However, these models focus on extracting features for a single task and fail to leverage the abundance of EEG data.

In this work, we propose an end-to-end self-supervised foundation model inspired by foundation models in speech, such as Wav2Vec [1] and WavLM [4], to learn a unified representation for EEG signals that can be used for multiple downstream tasks with minimal fine-tuning.

2 Related Work

Recent advancements in EEG signal processing have focused on improving the quality of raw EEG data by addressing the challenges posed by artifacts such as myogenic (muscle movement) and ocular (eye movement) noise, which can severely degrade the accuracy of EEG-based applications like Brain-Computer Interfaces (BCIs). [3] explored the use of Generative Adversarial Networks (GANs) to denoise EEG signals, leveraging both supervised and unsupervised learning. Their GAN-based framework outperformed traditional methods, improving Signal-to-Noise Ratio (SNR) and power spectral density (PSD), and demonstrated significant effectiveness in real-world applications by removing common EEG artifacts. [10] proposed an interpretable Convolutional Neural Network (CNN) model for detecting artifacts in single-channel EEG signals, with a focus on interpretability. The 1D-CNN model achieved high classification accuracy (around 99%) and effectively distinguished between ocular and myographic artifacts, making it suitable for real-time artifact detection in wearable EEG systems. [2] introduced the Riemannian Potato Field (RPF), a tool for assessing the Signal Quality Index (SQI) of EEG in real-time. By leveraging Riemannian geometry to compute distances between covariance matrices of EEG signals, the RPF demonstrated superior sensitivity and specificity in artifact detection, particularly in non-controlled environments. [11] compared the Artifact Subspace Reconstruction (ASR) method with traditional techniques like Independent Component Analysis (ICA) and Principal Component Analysis (PCA) for correcting EEG artifacts. The results highlighted ASR’s ability to outperform ICA and PCA, especially in preserving signal quality and correcting artifacts without significant signal loss, which is crucial for real-time applications such as BCIs. Finally, [16] proposed Deep Riemannian Networks (DRNs), specifically the EE(G)-SPDNet, which combines deep learning with Riemannian geometry to optimize custom filterbanks for EEG signals. This end-to-end model outperformed traditional CNN models in decoding motor imagery tasks, showing great promise for high-performance EEG decoding and real-time applications. These advancements demonstrate the growing role of deep learning, Riemannian geometry, and GANs in enhancing EEG signal quality and making EEG-based systems more viable for practical, real-time applications.

3 Datasets

3.1 Training Data

The Temple University Hospital EEG Corpus (TUH) [6] is a comprehensive and publicly available dataset of clinical EEG recordings. Collected over 14 years at Temple University Hospital, it comprises over 26,846 sessions from more than 10,000 unique subjects, making it one of the largest collections of clinical EEG data available for research. The dataset includes EEG recordings in the European Data Format (EDF) and corresponding clinician reports, providing valuable insights into patient conditions and clinical impressions. The TUH-EEG Corpus has been instrumental in advancing machine learning research in EEG analysis, particularly in areas like seizure detection and epilepsy diagnosis, by offering a diverse and extensive dataset that captures real-world variability in clinical settings.

3.2 Downstream Task Evaluation Data

We will be using the P300 BNCI2014-009 dataset to benchmark our foundation model against other baselines. P300, also known as an oddball paradigm, is a cognitive test used to measure brain activity related to attention, working memory, and decision-making. It involves presenting a participant with a series of stimuli, where a target stimulus is infrequently interspersed with more frequent non-target stimuli. The participant’s task is to detect and respond to the target stimuli. BNCI2014-009 consists of 10 subjects, each recorded for three sessions, and measured with 16 electrodes. Given the nature of the task the dataset is heavily skewed toward the non-target class with 80% of the samples being non-target and 20% belonging to the target class.

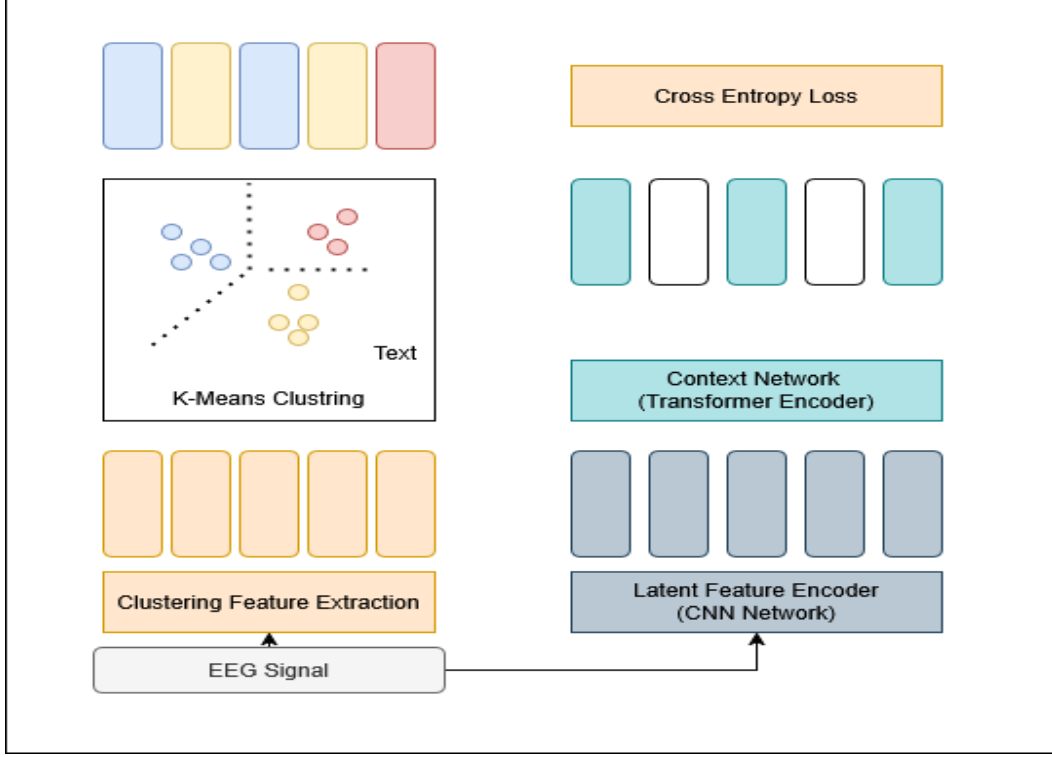


Figure 1: Training Process for HuBERT-EEG. (Grey tokens are masked at random). (White tokens are the predicted masked tokens from the previous step)

4 Methodology

4.1 HuBERT

The proposed EEG2VEC model is based on HuBERT [7] architecture, HuBERT is a self-supervised model that uses pseudo labels generated by a separate Kmeans model to perform masked token prediction loss, this paradigm allows HuBERT to learn meaningful information from speech signals that can be used in multiple downstream tasks such as Speech Emotion Recognition [9] and Speaker Verification [15]. By adapting this approach to EEG signals, EEG2VEC aims to leverage the strengths of HuBERT’s self-supervised learning framework to extract nuanced features from EEG data. This could potentially enhance the model’s ability to perform tasks like seizure detection, sleep stage classification, or cognitive state analysis, benefiting from the robust representations learned through masked token prediction. The use of pseudolabels allows the model to operate without explicit annotations, making it particularly suitable for large-scale EEG datasets where labeled data might be scarce. By extending HuBERT’s capabilities to EEG analysis, EEG2VEC offers a promising avenue for advancing our understanding of neural activity and improving diagnostic tools in neurology. In the context of EEG signals and HuBERT, a "token" refers to a segment or a feature extracted from the EEG time series. Unlike speech tokens, which are typically subwords or phonemes, EEG tokens are derived from the continuous signal by dividing it into segments or by extracting specific features such as spectral power or wavelet coefficients. These tokens serve as the basic units for the masked token prediction task, allowing the model to learn representations that capture the underlying patterns and structures within the EEG data.

4.2 Adapting HuBERT for EEG

HuBERT was originally designed to process speech, and as such most of its design choices are optimized for speech signals. This poses several challenges when applying it to EEG: unlike speech, EEG is inherently low in signal-to-noise ratio, is analyzed in short windows (0.2 ms to 3 s), and

is multichannel with a relatively low sampling rate (250 Hz in our case), whereas speech is mono-channel at 16 kHz.

To address these differences, we first configured the MFCC feature extractor for fixed 5 s chunks, specifying 64 MFCC bins, 128 FFT bins, a hop length of 64 samples, and 64 Mel bins. Next, we determined the number of clusters each representing a token of approximately 25 ms following HuBERT’s original design of learning granular features initially and then expanding token length to capture higher level abstractions; this choice yielded around 15 clusters. Finally, we scaled down the CNN encoder, which was originally sized for speech data, to suit the smaller EEG dataset.

Model	Hidden layers	Parameters/hyper-parameters
CNN1	(Conv1d, dropout, fpGroupNorm)	(1, 512, 125, 3), (0.0), (512, 512)
CNN2	(Conv1d, dropout)	(512, 512, 64, 1), (0.0)
proj1	(Linear, dropout-input, dropout-features)	(512, 768), (0.1), (0.1)
CNN3	(Conv1d)	(1, 512, 125, 3)
Transformer Encoderx12	(K, V, Q, out-proj)	(768, 768), (768, 768), (768, 768), (768, 768)
Self-attention	(LayerNorm, Linear, Linear, LayerNorm)	(768,), (768, 3072), (3072, 768), (768,)
layer-norm	(LayerNorm)	(768,)
layer-norm	(LayerNorm)	(512,)
final-proj	(Linear)	(768, 256)

Table 1: Architecture of HuBERT-EEG

4.3 Masked Token Prediction Loss

In the context of HuBERT, the masked token prediction loss is utilized to train the model in a self-supervised manner. Let $M \subset \{1, \dots, T\}$ denote the set of indices to be masked for a length- T sequence X . A corrupted version of X , denoted as $\tilde{X} = r(X, M)$, is created by replacing x_t with a mask embedding $\tilde{x}$ if $t \in M$. The masked prediction model f takes $\tilde{X}$ as input and predicts a distribution over the target indices at each timestep. This process allows the model to learn meaningful representations by predicting the original tokens from their masked versions, which is crucial for extracting robust features from speech signals.

Mathematically, this can be represented as follows:

Masking: For a sequence $X = (x_1, x_2, \dots, x_T)$, a subset of indices M is randomly selected to be masked. The corrupted sequence $\tilde{X}$ is then defined as:

$$\tilde{x}_t = \begin{cases} \tilde{x} & \text{if } t \in M \\ x_t & \text{otherwise} \end{cases}$$

Prediction: The model f predicts the original token x_t for each masked position $t \in M$, based on the context provided by $\tilde{X}$. This prediction is typically done by outputting a probability distribution over all possible tokens.

The masked token prediction loss is then computed as the cross-entropy between the predicted distribution and the true token at each masked position. This loss function encourages the model to learn representations that are informative about the original tokens, even when parts of the input are masked.

4.4 Evaluation Metric

To evaluate the quality of our representations, we will use improvements in downstream task performance as a measure of the extracted features’ quality. To standardize the evaluation process, we will utilize the SpeechBrain toolkit [13][12]. Although primarily designed for speech processing, the toolkit’s flexibility allows us to adapt it for EEG analysis. In this work, we concentrate solely on the P300 paradigm. By limiting our evaluation to P300 datasets, we can deeply analyze representation performance in this well-established ERP task.

The SpeechBrain toolkit offers several advantages for this evaluation:

- **Standardization:** It provides a consistent framework for evaluating EEG representations, ensuring that results are comparable across different models and datasets.
- **Task Focus:** Concentrating on P300 enables detailed investigation of feature quality in an ERP scenario where timing and amplitude differences are critical.

5 Baseline

5.1 EEGNet

EEGNet [8] is a compact convolutional neural network (CNN) designed specifically for EEG-based brain-computer interfaces (BCIs). It utilizes depthwise and separable convolutions to efficiently extract EEG features. EEGNet is notable for its ability to generalize across different BCI paradigms, such as P300, error-related negativity (ERN), movement-related cortical potentials (MRCP), and sensory motor rhythms (SMR), with high performance even when training data is limited. Its compact architecture allows it to achieve comparable results to more complex models while requiring significantly fewer parameters

5.2 Graph Neural Network

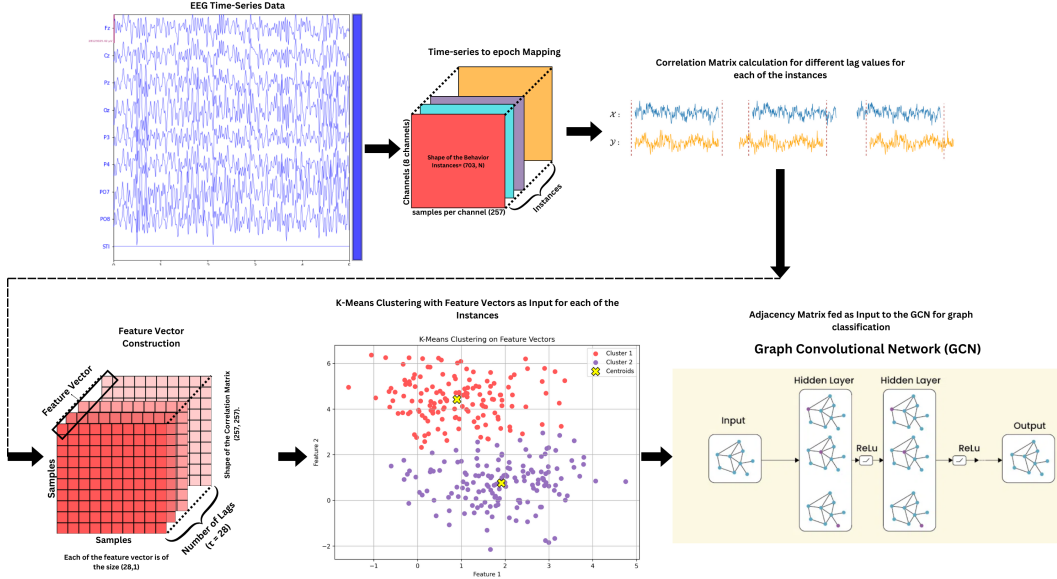


Figure 2: Framework for EEG-Graph Based Classification

Cross-Correlation and Feature Extraction: To capture functional and temporal relationships between EEG channels, we compute lagged cross-correlation matrices. For each EEG epoch $M \in \mathbb{R}^{C \times T}$ ($C = 8, T = 257$), lagged versions are generated via cyclic column shifts. For each lag $l \in \{0, \dots, 27\}$:

$$M_l = M \cdot C^l, \quad R^{(l)} = \frac{1}{C \cdot T} M M_l^\top$$

Flattening and stacking all $R^{(l)}$ yields a feature matrix $\mathbf{F} \in \mathbb{R}^{C^2 \times 28}$, encoding temporal dependencies for each channel pair.

Node Feature Computation: Node features are computed by averaging the original and lagged signals for each lag, then averaging across all lags:

$$F_{\text{node}} = \frac{1}{N} \sum_{l=0}^{N-1} \frac{1}{2} (M + M_l), \quad N = 28$$

Each channel’s row in $F_{\text{node}} (\in \mathbb{R}^{C \times T})$ serves as its node feature vector in the GCN.

Graph Construction: K-means clustering ($k = 2$) is applied to the channel-pair feature vectors to distinguish strongly and weakly correlated pairs. The resulting clusters define the adjacency matrix for the graph, representing the functional connectivity for each EEG epoch.

Graph Construction using K-means Clustering: To define the graph structure, the extracted feature vectors are clustered using K-means clustering with the number of clusters fixed at 2. This binary clustering identifies two types of connectivity:

- Strongly correlated channel pairs (connected)
- Weakly correlated or disconnected channel pairs (not connected)

An adjacency matrix is constructed from the clustering result, representing the graph topology for that EEG epoch.

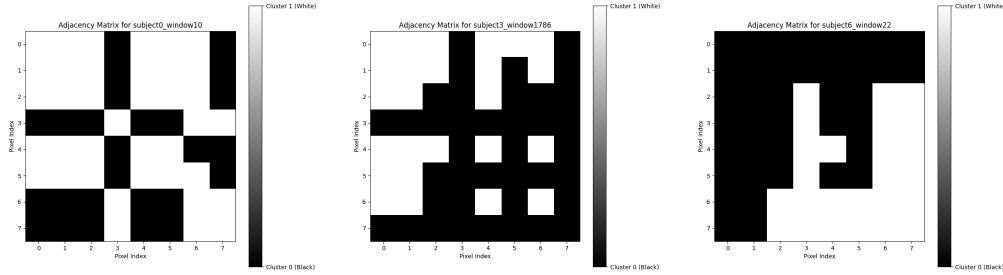


Figure 3: Samples of Adjacency Matrices

5.3 Separable Convolutional 1D Network

We employ a Separable Convolutional 1D (SepConv1D) model, designed to extract spatiotemporal patterns from time-series data. The model includes a separable convolution layer for efficient feature extraction, a Tanh activation, a flattening layer, and a fully connected sigmoid layer for binary classification. This architecture effectively captures both short-term and long-term dependencies in EEG signals.

5.4 Support Vector Machines

The raw EEG signals were first bandpass filtered between 5 Hz and 30 Hz using an 8th order Butterworth filter, then notch filtered at 60 Hz. To address per-subject class imbalance (3,500 nontarget vs. 700 target epochs), we augmented the minority class to achieve parity (3,500 nontarget and 3,500 target epochs). New target trials were generated by randomly sampling existing target epochs and applying three augmentation techniques—time masking, channel dropout, and Gaussian noise yielding 2,800 synthetic target epochs, which were then concatenated with the original dataset.

For each epoch and channel, we extracted seven time domain features and four frequency domain band powers:

- **Time domain features:** peak amplitude, signal area, mean, variance, skewness, kurtosis, and root mean square (RMS).
- **Frequency domain features:** band powers in delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), and beta (12–30 Hz) computed via Welch’s method.

These 11 features per channel produced an 88-dimensional feature vector for each epoch.

Baseline	ACC $\uparrow$
Graph Neural Network	77.93
Support Vector Machine	82.83
EEGNet	85.43
Separable Conv. 1D Network	87.14

Table 2: Baselines trained and evaluated on BNCI2014-009 P300 (80/20 split). 10 Subjects and 2 classes (target, non-target)

6 Results and Analysis

6.1 Baseline results

Each model is tested on the same individual’s data, leading to higher accuracy due to training on a homogeneous dataset. The mean accuracy achieved across subjects is 87.14% ($\pm 2.06\%$). The expected outcome of this project is to extract meaningful representations from EEG signals. The extracted representations should preserve the integrity of the underlying brain activity, which will then be fed into classification models or other task-specific algorithms. The goal is to improve the accuracy and robustness of EEG-based applications, such as Brain-Computer Interfaces (BCIs), cognitive state monitoring, or neurological disorder detection. In terms of performance, the project aims for an increase in Signal-to-Noise Ratio (SNR) and feature quality, leading to better classification results, faster processing times, and more reliable real-time performance.

Challenges include handling the variability and complexity of EEG signals across subjects and sessions, as well as ensuring that feature extraction retains important signal information. Risks involve potential overfitting to noise-free signals and balancing computational efficiency with high performance, especially in real-time applications.

6.2 HuBERT-EEG results

Model	ACC $\uparrow$
HuBERT-EEG	83.6
EEGNet	86.9

Table 3: Evaluated on BNCI2014-009 P300, trained on 8 subjects and evaluated on 2 unseen. 2 classes target and non-target

To perform downstream evaluation using HuBERT-EEG features, we employed a classification network inspired by EEGNet. As shown in our results, the performance of HuBERT-EEG lags behind that of the baseline model. Several factors may contribute to this outcome:

CNN feature extraction: Although we adapted the CNN encoder of HuBERT-EEG for EEG signal processing, there is still room for improvement in the feature extraction process. Incorporating spatial convolution could enable the model to better capture spatial relationships between EEG channels, while temporal convolution would help encode temporal dynamics across time. Additionally, introducing an attention mechanism to weigh the importance of each channel before passing the data to the CNN module could further enhance the network’s ability to focus on the most informative signals.

Training Data: While the TUH dataset provides a large repository of clinically recorded EEG signals, it does not necessarily include data corresponding to specific stimulus responses, such as P300 or motor imagery tasks. The lack of such targeted signals may limit the model’s ability to generalize to tasks that require detection of event-related potentials or other stimulus-specific patterns. To address this limitation, future work should incorporate datasets from a variety of experimental paradigms, thereby exposing the model to a broader range of neural responses.

Data Preprocessing: In our experiments, each EEG recording was segmented into fixed 5-second chunks. This approach poses several challenges: the model may not generalize well to shorter signals it has not encountered during training, and a fixed 5-second window may dilute the target signal or contain multiple overlapping events, depending on the downstream task [18]. As a result, the model’s ability to accurately detect and classify relevant neural events could be compromised. Moving

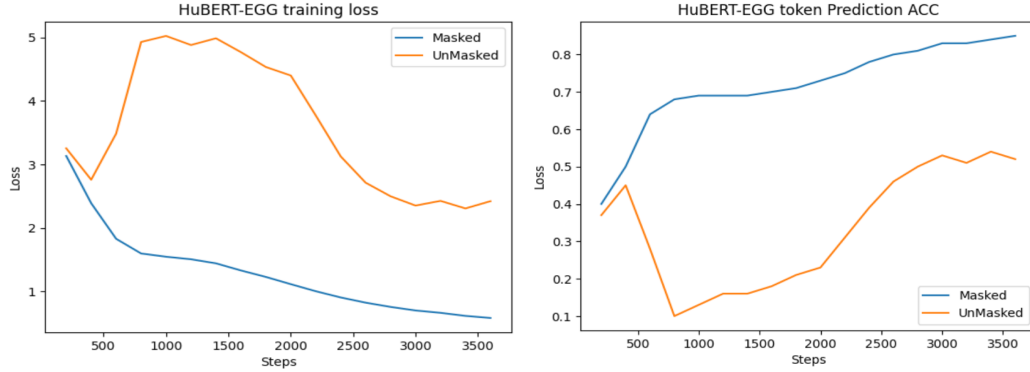


Figure 4: Loss and token prediction ACC during training

forward, we plan to segment recordings into variable-length chunks to better capture the diversity of signal durations and improve the model’s adaptability to different analysis windows.

Training Time: Due to the time required for experimentation with the CNN module, we were only able to train HuBERT-EEG for 79,000 steps out of a planned 400,000 steps. This incomplete training may have significantly impacted the model’s performance, as it did not have sufficient opportunity to fully learn from the data. The results presented in Table 3 likely reflect this limitation, and we anticipate that completing the full training schedule will lead to improved outcomes.

In summary, the current results highlight several areas for enhancement, including more sophisticated feature extraction, expanded and more diverse training data, improved data segmentation strategies, and extended training durations. Addressing these factors is expected to yield substantial improvements in HuBERT-EEG’s downstream performance.

7 Future Direction

The next steps for this project involve several key enhancements to improve the performance and generalizability of the proposed model. First, we plan to enhance the CNN module by incorporating both spatial and temporal convolutional layers. This will allow the network to more effectively capture spatial dependencies across channels as well as temporal patterns within the signal, thereby improving feature extraction.

Additionally, we aim to introduce stimulus-response datasets, such as those containing P300 event-related potentials, to expose the model to more diverse scenarios than those present in TUH. This will help generalize the capabilities of the model further across a wider range of brain-computer interface (BCI) tasks.

Another important direction is to move from using a fixed segment length to employing dynamic segment lengths. This adjustment will enable the model to better accommodate the inherent variability in neural signals and potentially improve classification accuracy by focusing on the most informative segments.

We also plan to experiment with various masking percentages and data augmentation techniques. By systematically varying the proportion of masked input and applying augmentation strategies, we hope to enhance the model’s robustness to noise and its ability to generalize to unseen data.

Finally, we propose to add an adaptive pooling layer before the CNN encoder. This layer will be designed to extract the most salient information from the input signal, ensuring that only the most relevant features are passed on for further processing. This step is expected to improve both the efficiency and effectiveness of the overall model.

Collectively, these future directions are intended to further advance the capabilities of our approach and broaden its applicability to diverse neural signal analysis tasks.

8 Conclusion

The results reported in this study, while not yet reaching the desired level of performance, underscore the significant challenges involved in developing a foundation model for EEG signals. These findings highlight the complexity of capturing the diverse temporal and spatial dynamics inherent in EEG data and suggest that further methodological advances are necessary.

Nevertheless, the feasibility of developing robust EEG foundation models is supported by recent successes such as EEGFormer, which has demonstrated promising results in this domain. These advances suggest that, with greater attention to factors like data diversity and architectural enhancements, further improvements in model performance are attainable.

It is important to emphasize that this work represents only the initial step in our ongoing project. Moving forward, we will conduct a series of experiments aimed at enhancing the quality of the representations learned by our model. Through iterative refinement and systematic evaluation, we aim to address the current limitations and advance toward a more effective and generalizable EEG foundation model.

9 Administrative Details

9.1 Division of work

Names sorted by last name:

Hazim Bukhari: Training and evaluation of HuBERT-EEG, and EEGNet baseline, pre-processing of TUH.

Mayur S Gowda: GNN, and Separable Convolutional 1D Network baselines and P300 dataset

Pranav Kini: Applying SVM baseline, Preprocessing and feature extraction of P300 dataset.

9.2 GitHub

<https://github.com/HazimBukhari/EEG2vec-A-unified-representation-for-EEG-signals>

References

- [1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- [2] Quentin Barthélemy, Louis Mayaud, David Ojeda, and Marco Congedo. The riemannian potato field: a tool for online signal quality index of eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(2):244–255, 2019.
- [3] Eoin Brophy, Peter Redmond, Andrew Fleury, Maarten De Vos, Geraldine Boylan, and Tomás Ward. Denoising eeg signals for real-world bci applications using gans. *Frontiers in Neuroergonomics*, 2:805573, 2022.
- [4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.
- [5] Mohammad-Javad Darvishi-Bayazi, Mohammad Sajjad Ghaemi, Timothee Lesort, Md Rifat Arefin, Jocelyn Faubert, and Irina Rish. Amplifying pathological detection in eeg signaling pathways through cross-dataset transfer learning. *Computers in Biology and Medicine*, 169:107893, 2024.
- [6] Amir Harati, Silvia Lopez, I Obeid, J Picone, MP Jacobson, and S Tobochnik. The tuh eeg corpus: A big data resource for automated eeg interpretation. In *2014 IEEE signal processing in medicine and biology symposium (SPMB)*, pages 1–5. IEEE, 2014.
- [7] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.
- [8] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.
- [9] Edmilson Morais, Ron Hoory, Weizhong Zhu, Itai Gat, Matheus Damasceno, and Hagai Aronowitz. Speech emotion recognition using self-supervised features. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6922–6926. IEEE, 2022.
- [10] Francesco Paissan, Velu Prabhakar Kumaravel, and Elisabetta Farella. Interpretable cnn for single-channel artifacts detection in raw eeg signals. In *2022 IEEE Sensors Applications Symposium (SAS)*, pages 1–6. IEEE, 2022.
- [11] Malgorzata Plechawska-Wojcik, Monika Kaczorowska, and Dariusz Zapala. The artifact subspace reconstruction (asr) for eeg signal correction. a comparative study. In Jerzy Świątek, Leszek Borzowski, and Zofia Wilimowska, editors, *Information Systems Architecture and Technology: Proceedings of 39th International Conference on Information Systems Architecture and Technology – ISAT 2018*, pages 125–135, Cham, 2019. Springer International Publishing.
- [12] Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, Francesco Paissan, Davide Borra, Salah Zaiem, Zeyu Zhao, Shucong Zhang, Georgios Karakasidis, Sung-Lin Yeh, Pierre Champion, Aku Rouhe, Rudolf Braun, Florian Mai, Juan Zuluaga-Gomez, Seyed Mahed Mousavi, Andreas Nautsch, Xuechen Liu, Sangeet Sagar, Jarod Duret, Salima Mdhaffar, Gaelle Laperriere, Mickael Rouvier, Renato De Mori, and Yannick Esteve. Open-source conversational ai with SpeechBrain 1.0, 2024.
- [13] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin,

- William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. SpeechBrain: A general-purpose speech toolkit, 2021. *arXiv:2106.04624*.
- [14] Siyi Tang, Jared A Dunnmon, Khaled Saab, Xuan Zhang, Qianying Huang, Florian Dubost, Daniel L Rubin, and Christopher Lee-Messer. Self-supervised graph neural networks for improved electroencephalographic seizure analysis. *arXiv preprint arXiv:2104.08336*, 2021.
 - [15] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*, 2021.
 - [16] Daniel Wilson, Robin Tibor Schirrmester, Lukas Alexander Wilhelm Gemein, and Tonio Ball. Deep riemannian networks for end-to-end eeg decoding, 2024.
 - [17] Ke Yi, Yansen Wang, Kan Ren, and Dongsheng Li. Learning topology-agnostic eeg representations with geometry-aware modeling. *Advances in Neural Information Processing Systems*, 36, 2024.
 - [18] Xinliang Zhou, Chenyu Liu, Zhisheng Chen, Kun Wang, Yi Ding, Ziyu Jia, and Qingsong Wen. Brain foundation models: A survey on advancements in neural signal processing and brain discovery. *arXiv preprint arXiv:2503.00580*, 2025.