Neural Decoding for Brain-Machine Interfaces: Vision Transformer and ECoG Feature Engineering

Théo Maëtz
Integrated Neurotechnologies Laboratory
EPFL Campus Biotech
Geneva, Switzerland
theo.maetz@epfl.ch

***Abstract*—Brain-Machine Interfaces (BMIs) are a transformative technology for restoring mobility in individuals with motor impairments. This study focuses on extending BMI applications to classify upperbody movements using electrocorticography (ECoG) signals. The research introduces a transformer-based neural decoding model designed to handle the high-dimensional nature of ECoG data effectively. A strong emphasis was placed on feature engineering, leveraging advanced spectral analysis techniques to optimize neural decoding. These contributions aim to enhance decoding performance and advance BMIs toward restoring movement for individuals with chronic paralysis.**

***Index Terms*—Neural Decoding, Transformers, Brain-machine Interface, ECoG, Wavelet transform**

I. Introduction

Recent breakthroughs in brain-machine interfaces (BMIs) have opened new possibilities for individuals with motor impairments, enabling direct communication between the brain and external devices. A recent landmark achievement in this field is the study "Walking naturally after spinal cord injury using a brain–spine interface" (Wagner et al., 2023) [1]. This work suggests the idea of a digital bridge demonstrating how cortical signals, once decoded into motor intentions in real time, can be routed to a targeted spinal cord stimulation system, effectively restoring a natural walking gait in patients with chronic paralysis.

Building upon this foundational work, our project focuses on extending to upperbody movement, exploring new possibilities for enhancing mobility and functionality in individuals with paralysis.

II. Related work

A. Brain–Machine Interfaces for Motor Rehabilitation

BMIs have evolved significantly over the past decades, offering potential solutions for individuals with severe motor impairments. BMIs rely on acquiring informative recordings of brain activity that can be accurately decoded to translate neural intentions into actionable commands. The use of electrocorticography (ECoG) as a reliable neural signal source for decoding motor intentions has been extensively validated in foundational studies [2] [3] [4] [5] [6].

A pivotal breakthrough was demonstrated by Shin et al. (2012) [7] and Sato et al. (2013) [8] by utilising ECoG recordings to decode motor intentions for prosthetic arm control in primates, emphasizing the viability of high-resolution spatial-temporal cortical signals for precise movement prediction. ECoGs provides a higher resolution with better signal-to-noise ratio than scalp electroencephalogram (EEG). ECoG has also shown potential as a stable long-term recording method.

A pivotal breakthrough in BMI research was demonstrated by Benabid et al. (2019) [9], where a tetraplegic patient controlled an exoskeleton through an epidural wireless BMI. This study showcased the ability of BMIs to provide functional mobility, albeit in controlled experimental settings, highlighting their potential for restoring motor functions in real-world applications.

B. ECoG Analysis

Wavelet transforms have emerged as a powerful tool for analyzing non-stationary neural signals like ECoGs [10] or even with spike signals for other neural decoding applications [11]. Wavelets provide a time-frequency representation that captures both transient and oscillatory features. This makes them particularly useful for decoding motor intentions from ECoG signals.

C. Advances in Neural Signal Decoding

Accurate decoding of neural signals is critical for an effective BMI implementation. The integration of high-gamma power features from ECoG recordings has proven particularly effective in classifying movement intentions. Ryun et al. (2017) [12] demonstrated that high-gamma power features from the human sensorimotor cortex can be used to accurately classify active movements. Paolo Viviani et al. (2023) [13] proposed a deep learning-based approach using Long Short-Term Memory (LSTM) networks to classify time-series neural data into grasp types. Although implemented without involving any prior neuroscience knowledge, the method demonstrates significant improvements in classification accuracy for real-time neural decoding applications. Biyan Zhou et al. (2023) [14] compared artificial neural networks (ANNs) and spiking neural networks (SNNs) for motor decoding in implantable BMIs. These works illustrate the potential of deep learning models in advancing the domain of neural decoding by enabling more accurate and robust interpretations of neural activity.

D. Transformers for Neural Decoding

Transformers provide a promising alternative for decoding complex and high-dimensional neural signals with their self-attention mechanisms. Their ability to model global dependencies and attend to key features across the entire input sequence makes them highly suitable for neural decoding tasks. Despite this potential, only limited research has explored the use of transformers in this domain [15] [16].

III. Methods

A. Objectives

This project builds upon the work presented in the study "Walking naturally after spinal cord injury using a brain–spine interface" (Wagner et al., 2023) [1]. The original research demonstrated the viability of a digital bridge linking cortical activity to spinal cord stimulation, enabling natural walking in individuals with chronic paralysis. While this achievement represents a major milestone in neurorehabilitation, further advancements are necessary to scale up the system for widespread practical use.

To address these challenges, our project focuses on developing a custom machine learning model tailored to the unique requirements of a BMI. We first design and train a model capable of accurately decoding ECoGs into movement intentions. Achieving high decoding accuracy is essential for ensuring the system's reliability and usability in real-world conditions. To evaluate the model's performance we will use the following metrics:

- Monitor the **accuracy** of the model as a general indicator of performance.
- Compute the **weighted F1 score,** ensuring balanced performance across all movement classes, even with imbalanced class distribution.
- **Averaged trace of the confusion matrix**, providing a summary of the correctly classified samples across all classes.

From this, we want to optimize the model for hardware efficiency, reducing the size and computational complexity of the model to ensure it is hardware-efficient. For this we will simply keep track of the model's number of parameters.

B. Data and Experimental Setup

The data used is the same as described in the original paper [1]. The ECoG data were collected from 32 channels per implant at an acquisition frequency of 586 Hz. Each data point consists of the ECoG signal of all 32 channels paired with the current movement state. Here we will use the ECoGs as our features $\mathbf{X}$ and the movement states as our labels $\mathbf{y}$. Assuming only a single movement is achieved at given instance we classify $\mathbf{y}$ over 6 possible movement classes. Therefore the raw data is $\mathbf{X}$ of shape (C, N) and $\mathbf{y}$ of shape $(1, N)$ with C the number of ECoG channels and N the number of data points.

The dataset used in this project was pre-split into training and testing sessions, with the split determined by the researchers who conducted the recording sessions with the patient. During these recording sessions, the `is_updating` flag was used to identify and retain only the data deemed relevant for decoding tasks.

C. Preprocessing of data

Preprocessing steps are necessary to ensure the quality and relevance of the neural signals used for decoding. One of the key components of the preprocessing pipeline is the MNE Filter, which leverages functions from the MNE-Python library for filtering ECoG data. This filter consists of bandpass filtering retaining only the frequency components between 1 and 200 Hz, notch filtering is applied at specified frequencies (50, 100, 150, and 200 Hz) as specified in the original paper [1] and we use a common average referencing (CAR) to help reduce spatially correlated noise and highlight local activity patterns

Decoding Model

The classification of movement intentions based on ECoG signals requires a model capable of extracting and interpreting both spatial and temporal features.

In this project we adapt the **Vision Transformer (ViT)** architecture proposed by Dosovitskiy et al. (2020) [17] to decode ECoG signals.

The ViT architecture, originally developed for image recognition tasks, provides a powerful framework for analyzing structured data, such as ECoG recordings, by treating them as sequences of patches. This transformer-based approach models relationships within the data using self-attention mechanisms, making it particularly effective for extracting both spatial and temporal features from ECoG signals. Additionally, its ability to generalize across high-dimensional inputs, further enhances its suitability for this application.

D. Data Windowing

To prepare the ECoG data for the model, we implement a sliding window approach that segments the continuous data into overlapping 3-second windows with a 1-second stride. This method ensures that each window captures sufficient temporal context for effective feature extraction and decoding.

The raw data X is windowed into a shape $(\cdot, C, 3 * T)$ and then further reshaped into $(\cdot, C * 3, T)$ with T being the number of time points of each window chosen to be the sampling frequency f_s. The second dimension $P = C * 3$ becomes the effective spatial dimension with each "spatial patches" now representing the signal from one channel over 3 seconds. The label y is also windowed accordingly and held at a constant value throughout the window.

E. Model Architecture

The model architecture used for this project is a ViT for ECoG-based neural decoding. An example of this model is given in figure 1. The backbone of this model is the transformer block which consists of L layers of Attention Blocks, which are themselves composed of two main components:

Multi-headed self-attention and a Feed-forward network with dropout layers, layer normalization, and GELU activation.

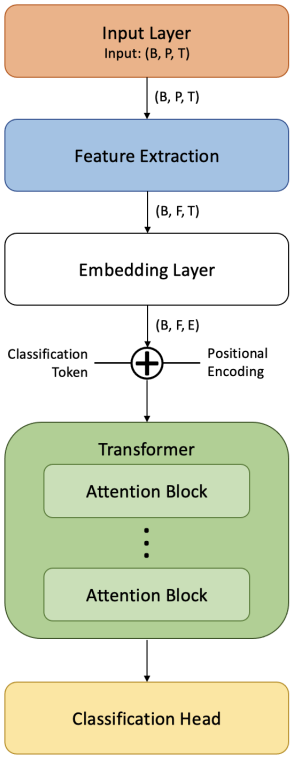

Fig. 1. Model architecture diagram, with B the batch size, P the number of patches, T number of time points, E the embedding dimension, F the feature dimension after feature extraction

F. Feature extraction

Feature extraction is a critical step in the model's architecture, as it transforms raw ECoG signals into a representation that is more meaningful and interpretable for the model. In this project, effective feature extraction is essential for capturing the spatial and temporal characteristics of the neural signals, which are key to decoding movement intentions.

Spectral features have also been shown to be highly informative for neural decoding [18] [19], resulting in improved decoding accuracy over traditional deep learning models that solely focus on spatio-temporal features.

We propose the use of spectrograms, similarly to the original paper [1], to simultaneously capture spatial, spectral, and temporal dynamics. Spectrograms allow us to visualize and extract the time-varying frequency content of ECoG signals. Given their effectiveness, our approach begins with a straightforward extraction of spectral features using the Fast Fourier Transform (FFT) on each patch of data.

We also want to leverage wavelet transforms to represent spectrograms. Starting with the Discrete Wavelet Transform (DWT), which is widely used in neural signal processing, providing a compact representation of localized frequency information, we can capture both time and frequency aspects effectively.

Finally, we explore the Continuous Wavelet Transform (CWT) to capture a richer representation of the signal. Unlike the discrete form, CWT offers a more detailed view of the time-frequency structure, making it particularly useful for analyzing non-stationary signals such as ECoG data.

G. Feature selection

While CWT provides a rich time-frequency representation of the signal, its granularity results in a significantly large number of features, which can increase computational complexity and the risk of overfitting. The high dimensionality also heavily increased the computational burden during training, especially for a complex model like the ViT. To address these challenges, we applied Principal Component Analysis (PCA) as a feature selection method.

IV. Results and Discussion

The model using FFT on each patch provides strong results to start with an accuracy of 72.7% and a weighted F1 score of 0.758, however the averged trace of the confusion matrix is at a relatively low 0.403.

A. DWT Results

DWT decomposes a signal into multiple levels, with each level representing a specific frequency band as shown with the example in figure 2.

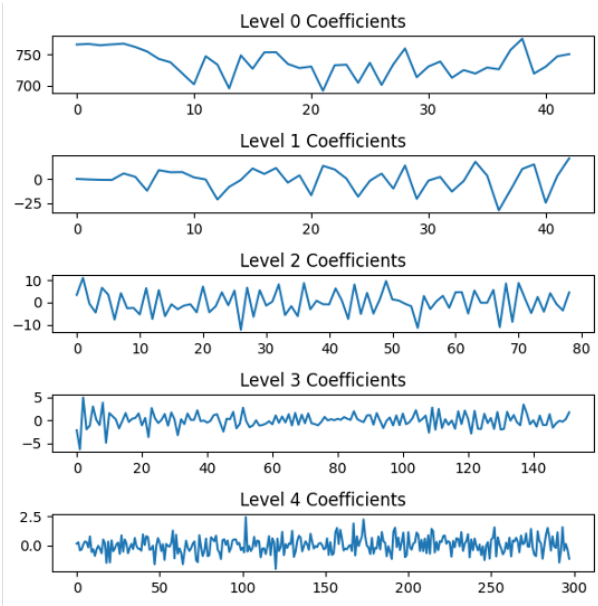

Fig. 2. Example of a 4 level DWT decomposition of a 1 second signal with a sampling frquency of 590 Hz

The initial DWT model simply concatenated its coefficients to form a feature vector. This concatenation across decomposition levels aggregates the frequency information without retaining the temporal context and also loses the hierarchical aspect of the decomposition. This issue could be remedied by concatenating across the time dimension which requires upsampling the higher levels of decomposition that are using

TABLE I
COMPARISON OF MODEL PERFORMANCE WITH THE INITIAL VISION TRANSFORMER IMPLEMENTATION

Model	Input Shape	Number of Params	Accuracy (%)	Weighted F1 Score	Avg. Trace
FFT	(B, 93, 590)	20.0 K	0.727	0.758	0.403
DWT concatenated coeffs	(B, 93, 615)	20.4 K	0.619	0.615	0.307
DWT upsampled	(B, 372, 298)	55.7 K	0.578	0.586	0.263
DWT time & position encoding	(B, 372, 298)	56.0 K	0.676	0.687	0.353
CWT Model (morl)	(B, 744, 590)	54.2 K	0.615	0.606	0.327

TABLE II
COMPARISON OF BSIT MODEL PERFORMANCE WITH DIFFERENT INPUT FEATURES

Model		Input Shape	Num Parameters	Accuracy (%)	Weighted F1 Score	Avg. Trace
Wavelet samples		(B, 10, 768)	52.7 K	0.770	0.769	0.692
CWT morl fully sampled	24 scales	(B, 590, 744)	52.7 K	0.590	0.566	0.396
CWT cmor1-1 fully sampled	24 scales	(B, 590, 744)	73.5 K	0.658	0.643	0.504
CWT cgau1 fully sampled	24 scales	(B, 590, 744)	73.5 K	0.527	0.497	0.326
Averaged CWT cmor1-1	24 scales	(B, 10, 744)	52.0 K	0.565	0.539	0.369
PCA CWT cmor1-1 fully sampled	24 scales	(B, 10, 384)	40.4 K	0.544	0.530	0.377
PCA Wavelet samples	24 scales	(B, 10, 384)	40.4 K	0.596	0.544	0.372

different time scales at each level. However this method resulted in worse performance across all three metrics, which could be explained by the introduction of artifacts from the upsampling that do not represent the true signal dynamics. We can also see that this method more than doubled the number of parameters.

Another method to address the loss of temporal information is a DWT model incorporating time embeddings alongside the existing positional embeddings which also now keep track of the levels of the decomposition. This solution showed improvement in all three performance metrics compared to the base DWT model but still less performant than the FFT model.

B. CWT Results

Despite its computational efficiency to capture general trends and oscillations, DWT lacks the precision needed to identify fine-grained temporal patterns within a signal. Unlike DWT, CWT provides a dense, continuous map of the signal's time-frequency as shown with the example in figure 3.

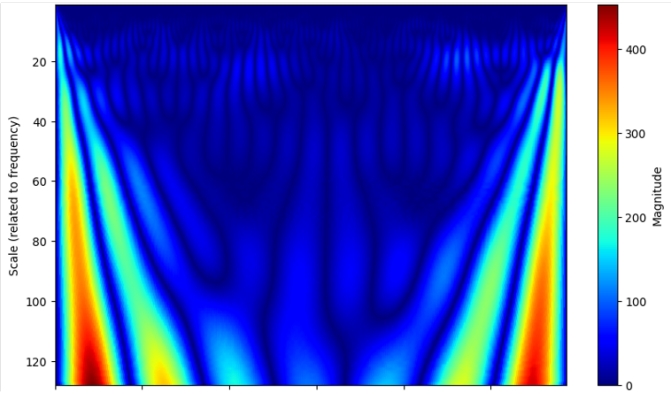

Fig. 3. Example of a 128 scale CWT Spectrogram of a signal

Furthering the CWT feature extraction

Going forward the following results were obtained using a provided model (BSIT) with slight differences from the original while still keeping the same main components of the model architecture shown in figure 1 however this time around using 1 second windows with a 0.1 second stride. This model obtained the best results when using the pre-computed wavelet features as described in the original paper [1].

The reason for continuing with this new model is its suitability for focusing on feature engineering and exploring further into feature extraction techniques. This model allows us to either replicate the results achieved with the wavelet samples or potentially achieve improved performance by leveraging the capabilities of the ViT. By concentrating on optimizing the input features, we aim to fully utilize the ViT's capabilities regarding neural decoding.

With this transition to a new model, we also opted to train on a single training session from the dataset to simplify the training process and focus on refining the model's architecture and feature extraction capabilities. Consequently, the movement intentions recorded within this specific session were limited to fewer distinct classes.

We began by evaluating the performance of different wavelet families using fully sampled CWTs: Morlet (morl), Complex Morlet (cmor), Complex Gaussian (cgau). Among these, the Complex Morlet wavelets yielded the best results. This result aligns itself with the provided wavelet samples that were also extracted using complex Morlet wavelets. A complex wavelet consists of real and imaginary components, capturing both magnitude and phase information providing a comprehensive representation of the signal which can improve classification performance.

C. Downsampling

Despite its potential drawbacks, downsampling could prove beneficial with regards to the high-dimensionality of neural signals, to reduce possible overfitting where the model learns

noise or irrelevant patterns instead of generalizable features. After identifying cmor as the most effective wavelet, we downsampled the fully sampled CWTs to match the format of pre-computed wavelet samples. The wavelet sample features were structured into a shape of $(10,\ 32 * 24)$ where 10 is the number of time windows and $32*24$ represent the 32 channels and the 24 scales used for the CWT.

The initial downsampling appproach consisted of averaging the signals over 0.1 second windows, which did cause a noticeable drop in performance.

D. Feature Selection

Instead of downsampling in time, we shifted focus to feature selection across the combined (channel $*$ scales) dimension for each time window. This approach aims to retain the most informative features (ECoG channels and frequency components) while reducing dimensionality.

Using this method we do get similar results as the averaging method, with a worse accuracy but a better averaged trace of the confusion matrix.

V. Conclusions

This work extends the potential of BMIs by focusing on upperbody movement classification. Building upon the foundational work of Wagner et al. [1], we used advanced feature extraction techniques and adapted the transformer-based architecture to decode ECoG signals. The integration of wavelet-based features, both discrete and continuous wavelet transforms, demonstrated the importance of rich time-frequency representations for neural decoding tasks. One aspect that wasn't explored regarding the use of the wavelet transforms are the issues with edge effects. The edge effects may distort the extracted features leading to inaccuracies in subsequent analysis.

Our experimentation revealed the strengths and limitations of various feature extraction and selection methods, the benefits of incorporating spectral features and the challenges of dimensionality reduction. Additionally, the adaptability of transformer architectures, like the ViT, for neural signal decoding was explored, and demonstrated promising results.

Further work

Building on the progress made in this study, there are several directions for future exploration and refinement. Promising results were achieved by applying the FFT to each time window of the ECoG data. Future work could explore further into refining short-time FFT (STFT) to capture temporal variations in the frequency domain to provide a spectrogram representation of the data.

The results with the wavelet features could be furher improved by remedying the issues relating to edge effects. We could add padding to each wiondow when applibg the wavelet transformation and only extract the center features of each padded windows.

In this study, while significant strides were made in optimizing the model architecture and feature extraction methods, a more exhaustive hyperparameter search remains an area for further exploration. Hyperparameters such as the number of transformer layers, attention heads, embedding dimensions, and learning rates play a crucial role in the model's performance.

To ensure real-time applicability, more emphasis should also be put on effective feature selection and dimensionality reduction techniques. Next step is translating the refined models and feature extraction pipeline into a real-time implementation.

Acknowledgements

I would like to express my heartfelt gratitude to my project supervisor Dr. Mohammad Ali Shaeri for entrusting me with this significant project, providing invaluable guidance throughout its development. I extend my deepest thanks to Arshia Azfal for their instrumental support in providing the initial setup for this project and the original model architecture. I am equally grateful to Yuhan Xie, who provided the BSIT model and whose continued assistance and expertise greatly contributed to the progress of this work.

I would also like to thank Professor Mahsa Shoaran of the Integrated Neurotechnologies Laboratory, for accepting me to work within her esteemed laboratory.

Lastly, I am immensely grateful to the Lighthouse Partnership for their commitment to restoring movement for patients with chronic paralysis. This project is a humble contribution to their overarching goal, made possible by the datasets they generously provided for our research.

References

[1] H. Lorach, A. Galvez, V. Spagnolo, F. Martel, S. Karakas, N. Intering, M. Vat, O. Faivre, C. Harte, S. Komi, J. Ravier, T. Collin, L. Coquoz, I. Sakr, E. Baaklini, S. D. Hernandez-Charpak, G. Dumont, R. Buschman, N. Buse, T. Denison, I. Van Nes, L. Asboth, A. Watrin, L. Struber, F. Sauter-Starace, L. Langar, V. Auboiroux, S. Carda, S. Chabardes, T. Aksenova, R. Demesmaeker, G. Charvet, J. Bloch, and G. Courtine, "Walking naturally after spinal cord injury using a brain–spine interface," *Nature*, vol. 618, no. 7963, p. 126–133, Jun. 2023.

[2] T. Pistohl, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring, "Prediction of arm movement trajectories from ecog-recordings in humans," *Journal of Neuroscience Methods*, vol. 167, no. 1, p. 105–114, Jan. 2008.

[3] T. Ball, M. P. Nawrot, T. Pistohl, A. Aertsen, A. Schulze-Bonhage, and C. Mehring, "Towards an implantable brain-machine interface based on epicortical field potentials."

[4] B. Pesaran, J. Pezaris, M. Sahani, P. Mitra, and R. Andersen, "Temporal structure in neuronal activity during working memory in macaque parietal cortex," *Nature Neuroscience*, vol. 5, no. 8, p. 805–811, 2002.

[5] G. Schalk, J. Kubánek, K. J. Miller, N. R. Anderson, E. C. Leuthardt, J. G. Ojemann, D. Limbrick, D. Moran, L. A. Gerhardt, and J. R. Wolpaw, "Decoding two-dimensional movement trajectories using electrocorticographic signals in humans," *Journal of Neural Engineering*, vol. 4, no. 3, p. 264–275, Sep. 2007.

[6] G. Schalk and E. C. Leuthardt, "Brain-computer interfaces using electrocorticographic signals," *IEEE Reviews in Biomedical Engineering*, vol. 4, p. 140–154, 2011.

[7] D. Shin, H. Watanabe, H. Kambara, A. Nambu, T. Isa, Y. Nishimura, and Y. Koike, "Prediction of muscle activities from electrocorticograms in primary motor cortex of primates," *PLoS ONE*, vol. 7, no. 10, p. e47992, Oct. 2012.

[8] K. Sato, S. Morishita, H. Watanabe, Y. Nishimura, R. Kato, T. Isa, and H. Yokoi, "Discrimination analysis and movement decision of the prosthetic arm by using monkey ecogs data associated with self-feeding motions," *Journal of the Robotics Society of Japan*, vol. 31, no. 1, p. 51–59, 2013.

[9] A. L. Benabid, T. Costecalde, A. Eliseyev, G. Charvet, A. Verney, S. Karakas, M. Foerster, A. Lambert, B. Morinière, N. Abroug, M.-C. Schaeffer, A. Moly, F. Sauter-Starace, D. Ratel, C. Moro, N. Torres-Martinez, L. Langar, M. Oddoux, M. Polosan, S. Pezzani, V. Auboiroux, T. Aksenova, C. Mestais, and S. Chabardes, "An exoskeleton controlled by an epidural wireless brain–machine interface in a tetraplegic patient: a proof-of-concept demonstration," *The Lancet Neurology*, vol. 18, no. 12, p. 1112–1122, Dec. 2019.

[10] A. Runnova, M. Zhuravlev, R. Ukolov, I. Blokhina, A. Dubrovski, N. Lezhnev, E. Sitnikova, E. Saranceva, A. Kiselev, A. Karavaev, A. Selskii, O. Semyachkina-Glushkovskaya, T. Penzel, and J. Kurths, "Modified wavelet analysis of ecog-pattern as promising tool for detection of the blood–brain barrier leakage," *Scientific Reports*, vol. 11, no. 1, p. 18505, Sep. 2021.

[11] A. Soleymankhani and V. Shalchyan, "A new spike sorting algorithm based on continuous wavelet transform and investigating its effect on improving neural decoding accuracy," *Neuroscience*, vol. 468, p. 139–148, Aug. 2021.

[12] S. Ryun, J. S. Kim, E. Jeon, and C. K. Chung, "Movement classification using ecog high-gamma powers from human sensorimotor area during active movement," in *2017 5th International Winter Conference on Brain-Computer Interface (BCI)*. Gangwon Province, South Korea: IEEE, Jan. 2017, p. 96–98. [Online]. Available: http://ieeexplore.ieee.org/document/7858171/

[13] P. Viviani, I. Gesmundo, E. Ghinato, A. Agudelo-Toro, C. Vercellino, G. Vitali, L. Bergamasco, A. Scionti, M. Ghislieri, V. Agostini, O. Terzo, and H. Scherberger, "Deep learning for real-time neural decoding of grasp," 2023. [Online]. Available: https://arxiv.org/abs/2311.01061

[14] B. Zhou, P.-S. V. Sun, and A. Basu, "Ann vs snn: A case study for neural decoding in implantable brain-machine interfaces," 2023. [Online]. Available: https://arxiv.org/abs/2312.15889

[15] R. Liu, M. Azabou, M. Dabagia, J. Xiao, and E. L. Dyer, "Seeing the forest and the tree: Building representations of both individual and collective dynamics with transformers," 2022. [Online]. Available: https://arxiv.org/abs/2206.06131

[16] I. Han, J. Lee, and J. C. Ye, "Mindformer: Semantic alignment of multi-subject fmri for brain decoding," 2024. [Online]. Available: https://arxiv.org/abs/2405.17720

[17] 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," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[18] K. G. Hartmann, R. T. Schirrmeister, and T. Ball, "Hierarchical internal representation of spectral features in deep convolutional networks trained for eeg decoding," 2017. [Online]. Available: https://arxiv.org/abs/1711.07792

[19] X. Li, Y. Chu, and X. Wu, "3d convolutional neural network based on spatial-spectral feature pictures learning for decoding motor imagery eeg signal," *Frontiers in Neurorobotics*, vol. 18, p. 1485640, Dec. 2024.