Learning Representations from EEG Brain Signals

Hamza Amrani

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Milano-Bicocca

2025

Program Authorized to Offer Degree:
Computer Science

Learning Representations from EEG Brain Signals

Abstract

Brain-computer interfaces (BCIs) promise direct communication between neural activity and external devices. Electroencephalography (EEG) offers a non-invasive approach through lightweight, portable hardware. Despite decades of research, however, EEG-based BCIs remain largely confined to laboratories. The core challenge lies not only in hardware limitations but also in extracting robust and meaningful representations from noisy and variable neural signals. Traditional pipelines extract task-specific features that work in controlled settings but not in new situations involving different users, sessions, or tasks. Performance drops by 19% when models trained on one subject are tested on another. This prevents the practical deployment of BCIs in assistive technologies, healthcare, and human-computer interaction.

This dissertation demonstrates how deep representation learning can overcome limitations by discovering invariant features directly from data and aligning neural patterns with human-interpretable semantics. Instead of engineering features for each application, we learn general-purpose representations that transfer across subjects, sessions, and tasks. We make three primary contributions: (1) Novel neural architectures including frequency-aware multi-band encoders and subject-conditional transformers that capture both neural patterns and individual differences; (2) One of the first end-to-end systems for open-vocabulary semantic decoding from EEG, achieving BLEU-1 of 42.75% and BERTScore-F of 53.86% by aligning brain representations with language model spaces across 30 subjects; (3) Neural tokenization via vector quantization that converts continuous signals into discrete

”brain tokens,” enabling efficient downstream models suitable for real-time, edge-deployed applications.

The methods have been validated across diverse experimental paradigms and hardware configurations, including consumer-grade dry-electrode devices and high-density clinical systems. This demonstrates the methods' practical viability across different acquisition settings. Comprehensive evaluations of motor imagery, emotion recognition, and semantic decoding tasks show that learned representations outperform handcrafted features and generalize across the variability that has historically limited BCI deployment. This work establishes representation learning as a viable path toward generalizable, interpretable, and deployable BCIs.

Acknowledgments

I want to express my sincere gratitude to the University of Milan-Bicocca for giving me the space and resources to explore unconventional paths and follow my curiosity wherever it led. My deepest thanks go to Carnegie Mellon University for welcoming me during a transformative year that broadened my perspective and enriched my journey, in ways I could not have anticipated.

I want to thank my advisors, Daniela Micucci and Paolo Napoletano, for teaching me not just how to conduct research, but how to think like a researcher—to question deeply, to navigate uncertainty, and to trust the process even when the destination wasn't clear. Their patience, mentorship, and willingness to entertain my more audacious proposals made this work possible.

I am especially grateful to Vittorio Caggiano and Vikash Kumar for hosting me during my research stays in Pittsburgh and New York. They challenged my ideas, welcomed me into their labs, and made those months abroad among the most formative of my PhD.

Finally, to everyone who walked alongside me through this journey. Your encouragement, support, and friendship turned years of hard work into cherished memories. Thank you.

Collaborators: Marco Mobilio, Ilario Rizzi, Luigi Ferrara, Alessandro Fornaro, Marco Nalin.

Institutions: Department of Informatics, Systems and Communication; Intelligent Sensing Laboratory; Software Architecture Laboratory; Imaging and Vision Laboratory; Robotics Institute; abmedica spa; myolab AI.

dedication

to my parents, Rajae and Mohamed

for believing in me every step of the way

to my fiancé, Lucia

for your love, patience, and support

to my brother, Adam, and my sisters, Nadia and Sofia

for always being there

You are the foundation upon which this achievement stands.

Contents

Listing of figures

Part I

Foundations

Everything we do, every thought we've ever had, is produced by the human brain. But exactly how it operates remains one of the biggest unsolved mysteries, and it seems the more we probe its secrets, the more surprises we find.

Neil deGrasse Tyson

1

Introduction

1.1 Learning from EEG Signals

The human brain generates electrical activity that encodes our thoughts, intentions, and perceptions. Measuring and interpreting these signals remains one of neuroscience's central challenges. Every second, billions of neurons fire in coordinated patterns, creating electromagnetic fields detectable through electroencephalography (EEG) and magnetoencephalography (MEG). Yet extracting meaningful information from these measurements proves remarkably difficult.

Researchers for many decades have relied on manual feature extraction in order to make sense of brain signals. An ordinary brain-computer interface (BCI) could register voltage differences in as few as 64 channels of electrodes, then manually compute pre-specified features such as power in user-selected frequency bands or the amplitude of event-related potentials. This process works well for simple tasks but fails to capture the full richness of information inherent in neural activity. The problem is more than a technical limitation—it is a fundamental mismatch between the complexity of brain signals and the analysis methods we design by hand.

Representation learning offers a different solution. Rather than specifying features manually, we can now learn models to discover optimal representations from data in an end-to-end approach (see Figure 1.1). This change in viewpoint, facilitated by advances in computer vision and natural language processing over the past ten years, could similarly reveal the patterns of human cognition in the nervous system. The question is not whether information can be decoded from brain signals — the answer is an emphatic yes — but how we can learn representations that encode that information effectively.

1.1.1 The Evolution of Neural Representation Paradigms

Research into neural representations has advanced through the use of different paradigms. Early neuroscience adopted a localist approach, whereby single neurons were believed to represent specific concepts, a theory known as the 'grandmother cell' hypothesis. However, experimental data quickly dis-

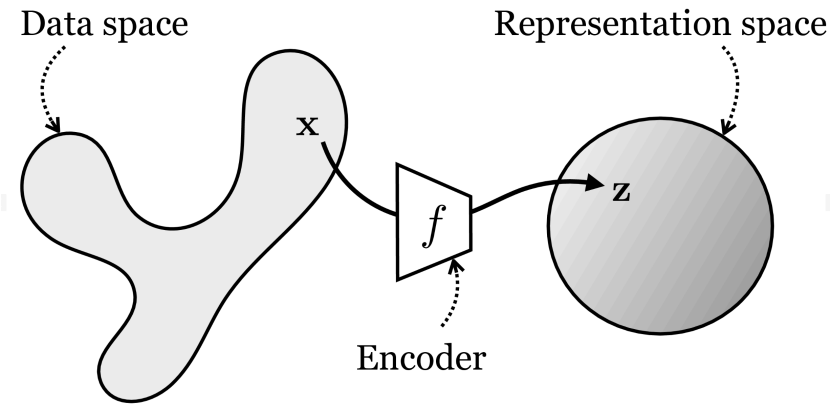


Figure 1.1: From data space to representation space. An encoder f maps an observation x to a representation z that is easier to separate, compare, and reuse across tasks.

proved this model as being too simplistic. Cells operate in populations, with information spread out across many units.

Hinton's 1984 paper on distributed representations set out the theory of population codes Hinton (1984). According to this theory, concepts emerge from distributed patterns of activation in large numbers of units rather than in single neurons. This was consistent with neuroscience findings indicating that motor commands, sensory perceptions, and memories all utilized distributed neural activation. The parallel distributed processing architecture showed that networks could learn such representations by varying connection weights based on error-driven procedures.

By 2013, Bengio et al. had formalized the principles of representation learning: models must learn features automatically rather than through human engineering Bengio et al. (2013). Deep networks learn hierarchical representations, with lower levels picking up simple properties and higher levels encoding abstract ideas. This hierarchical structure is similar to that of sensory cortices in the brain: neurons in the earliest visual regions respond to edges, while higher regions fire in response to faces or objects.

The emergence of self-supervised learning in the mid-2010s was of special significance. Instead of

being given labeled data for each example, models could learn from pretext tasks, such as predicting missing components of their input or distinguishing real from augmented data. Word2vec Mikolov et al. (2013) demonstrated that predicting nearby words generates meaningful word representations. In computer vision, models learned to create contextual representations by predicting image rotations or completing puzzles. This paradigm was particularly valuable for applications where labels were costly, such as with brain signals. Models could learn from the structural information in raw data without any manual annotation.

Around 2018, the age of foundation models began with GPT Brown et al. (2020) and BERT Devlin et al. (2019). Both utilized self-supervised training on a large scale, pretraining hundreds of billions of tokens to learn general-purpose representations. This made it possible to fine-tune a single model, trained once on mixed data, for tasks such as translation, question answering, and sentiment analysis with minimal task-specific training. Pre-training followed by fine-tuning became the default machine learning procedure. The effectiveness of this strategy raised the question of whether brain signals would benefit from it as well.

1.1.2 Challenges of Traditional EEG Analysis

Brain signals present unique challenges compared to images or text. EEG recordings typically include 64 to 256 channels, each of which samples at 250 to 1,000 hertz (Hz). A one-hour recording generates millions of data points. However, high dimensionality is only part of the challenge. The signals are non-stationary, meaning their statistical properties change over time. A person's mental state, level of fatigue, and level of attention all affect the recorded patterns. Even within a single recording session, signal characteristics change. Individual differences compound these challenges. The way electrical fields propagate through tissue is affected by brain anatomy, which varies from person to person. Skull thickness, cerebrospinal fluid volume, and cortical folding patterns also influence the signals recorded at the scalp. Two people performing identical mental tasks produce different EEG patterns. What

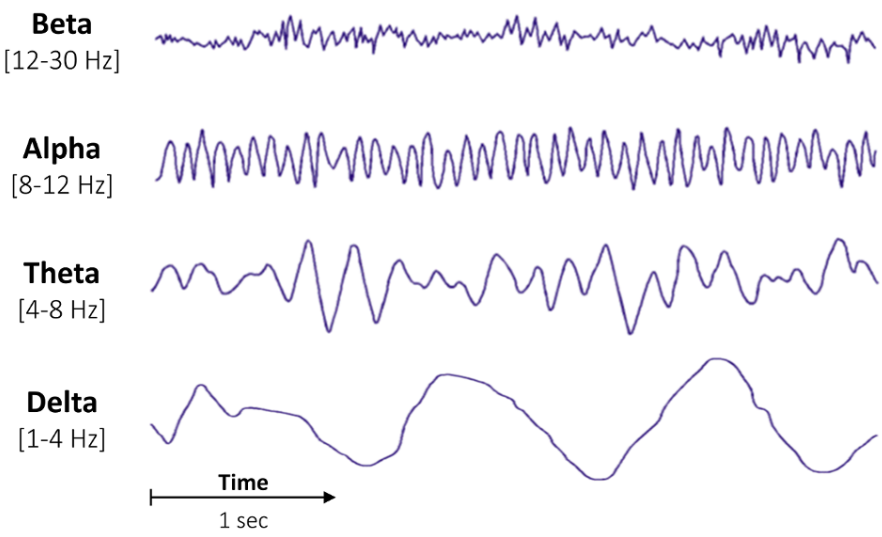


Figure 1.2: EEG data is to decompose the signal into functionally distinct frequency bands, such as delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–100 Hz).

works for one subject often fails for another. The signal-to-noise ratio in EEG is particularly poor, often around -10 dB. Electrical activity from deep brain structures is significantly reduced before it reaches scalp electrodes. Meanwhile, additional noise is introduced by muscle movements, eye blinks, and external electrical interference. The signals we want to decode are buried in artifacts that are orders of magnitude larger.

Additionally, traditional EEG systems extract features based on neuroscientific principles. Power spectral density captures the energy in different frequency bands (see Figure 1.2). For example, alpha waves are associated with relaxation, beta waves with active thinking, and gamma waves with attention. Event-related potentials (ERPs) measure voltage changes time-locked to stimuli. The P300 response, for example, appears 300 milliseconds after rare events. Common spatial patterns identify linear combinations of channels that maximize the variance between mental states. These features can be generalized to restricted, well-constrained tasks. For instance, to distinguish between visualizing moving your left hand versus your right hand, you can evaluate beta-band power in the motor

cortex. To detect attention, look for alpha power. However, for each task, features must be individually trained. Features that generalize for motor imagery fail to generalize for emotion recognition, and these in turn fail to generalize for visual perception tasks. Even more basic, manually crafted features fail to capture unforeseen data patterns. If neural information is carried in phase relations between distant brain sites, then spectral power features will miss it. If cognitive states comprise intricate spatiotemporal patterns extending beyond one or more of these bands, they cannot be represented by simple features.

Furthermore, EEG systems require lengthy calibration periods for each user. The process typically begins with the collection of 20-30 minutes of training data. During this time, the user performs defined mental tasks repeatedly while the system collects their brain signals. Machine learning models learn from this type of subject-dependent data. However, the calibration must be repeated in each session because the positions of the electrodes change and the properties of the signals fluctuate. This approach is not scalable. Clinical applications require real-time BCIs for new patients. Consumer products cannot require a half-hour preparation protocol. Calibration overhead also limits research; instructions must restrict sample sizes because each subject requires personalized training for their model. The crucial issue is that traditional models learn subject- and session-dependent regularities rather than invariant representations. When trained on one subject's data, a model's performance is random when tested on another subject. Even for the same participant, models trained on Monday would not generalize to Tuesday. Without representations that generalize from subject to subject and from session to session, BCIs will remain difficult to generalize.

1.1.3 Principles of Representation Learning for Brain Signals

Representation learning transforms the fundamental approach to brain signal analysis. Rather than asking, "What features should we extract?" we ask, "What representations do the data support?" The model discovers features through optimization rather than human design. Three key principles are

required for this. First, models must learn invariances rather than memorize specific patterns. A good representation captures commonalities across subjects while ignoring variations that do not affect the task. Second, unsupervised and self-supervised learning are essential. Collecting labeled data for brain signals is too costly, so we cannot rely solely on it. Models must learn from the structure in raw recordings. Third, learned representations should be hierarchical, building complex concepts from simpler components and matching the organization of the brain.

The next frontier—and the focus of this dissertation—is connecting these learned neural representations to human-interpretable semantic spaces, enabling true brain-to-world communication.

1.2 Bridging Neural Activity and Semantic Meaning

1.2.1 The Semantic Gap in Brain–Computer Interfaces

Representation learning overcomes the limitations of feature engineering by identifying ideal features within data. However, a much deeper challenge remains: closing the gap between semantic content and neural measurement. Although state-of-the-art brain-computer interfaces can sort signals into separate classes, distinguishing between left versus right hand motor imagery or positive versus negative affective states, they are unable to accurately decipher a person's real thoughts. Thus, the focus shifts from which task a person is performing to what semantic content their neural activity actually represents.

This challenge defines the semantic gap, which is the difference between bottom-up neural activity and top-down meaning. Most BCIs operate within a classification framework with preestablished labels. P300 spellers categorize which of 36 characters were recognized Farwell & Donchin (1988). Motor imagery interfaces differentiate between two and four mental states. Advanced emotion recognition translates affective experience into discrete classes or low-dimensional manifolds. Semantic decoding, by contrast, aims to retrieve meaningful content, such as words and sentences in language,

object identities and relations in vision, continuous motor parameters in motor control, and specific emotional concepts with contextual details.

The technical reason for this gap is the EEG's inherent limitations. Scalp measurements sample the activity of millions of neurons that are spatially smeared by volume conduction Cohen (2014). With spatial resolution constrained to 1-3 cm, 64-256 channels must represent thousands of potential semantic ideas. The inverse problem—finding the neural sources for scalp measurements—has infinitely many solutions. Signal-to-noise levels of -10 dB frequently result in artifacts that overwhelm the neural signals by a factor of 10.

Scientific uncertainty multiplies these technical constraints. While fMRI demonstrates systematic semantic structure in cortical activity, such as topographical maps for object classes, action categories, and abstract concepts Huth et al. (2016); Mitchell et al. (2008), it is unclear whether such structure is preserved in EEG due to its spatial resolution. Nevertheless, substantial evidence suggests that semantic information persists at the level of the scalp. The N400 event-related potential systematically indexes semantic violations and contextual unpredictability Kutas & Federmeier (2011). Multivariate pattern analysis can extract object category decoding, word identity decoding, and semantic similarity structures from EEG with above-chance accuracy King & Dehaene (2014). The millisecond time course of an EEG can detect rapid changes in language processing that an fMRI cannot.

The semantic gap is a significant yet manageable barrier. The hypothesis that the EEG signal contains semantic information is supported by empirical evidence. The key question is the construction of representations and structures that are capable of abstracting this information reliably. This abstracting must go beyond simple classification and address the time-continual, high-dimensional nature of the semantic content.

1.2.2 Approaches for Neural–Semantic Alignment

Two computational approaches have been explored to bridge the semantic gap. The first reconstructs semantic content directly from brain signals. The second learns shared embedding spaces between neural patterns and semantic content, such that the latter can be aligned with the former. Pre-trained foundation models are being repurposed to accept brain signals as just another modality of data.

Kay et al. (2008) were the first to show visual reconstruction by outlining how fMRI voxel patterns could be traced back to natural images. Recent research goes even further by pipelining brain activity through generative models. Takagi & Nishimoto (2023) map fMRI signals onto diffusion model latent space and thereafter employ the models to generate very accurate images of what subjects perceived. For language, Tang et al. (2023) aligned fMRI recordings against GPT internal representations with continuous decoding of spoken words. EEG makes it harder—the spatial blur caused by volume conduction erases fine-grained information. However, Défossez et al. (2023a) decoded speech from non-invasive recordings by synchronizing them with wav2vec features. The same has been tried with reading, but the success is not quite as great.

Contrastive methods work differently. Borrowing ideas from CLIP Radford et al. (2021), they build joint spaces where responses to a given stimulus lie near the semantic representation of the stimulus and responses to other stimuli are far away from each other. Training maximizes similarity for correct pairings and minimizes it for incorrect ones, typically through InfoNCE or triplet losses. This sidesteps the reconstruction problem entirely—you care about capturing meaning, not regenerating exact pixels or words. It also handles a thorny issue: different brain patterns can correspond to semantically similar concepts, and contrastive learning naturally accommodates this many-to-one mapping. Brain-Score work Schrimpf et al. (2018) showed that if you align deep network representations with actual neural recordings, the networks can make amazingly accurate predictions of brain activity patterns.

Recent work merges brain signals with existing foundation models. Language and vision models trained on large datasets already possess rich semantic structure. The real challenge is taking brain signals—high-dimensional, continuous—and making something transformers can use. Transformers process discrete tokens. Three solutions have been proposed. Adapter networks add small modules as a way of mapping brain representations to the input space of the model. Soft prompts convert brain information into continuous vectors that are added as a prefix to sequences. Tokenization converts brain signals into vocabularies, which are processed as words. This last approach enables brain signals to act as genuine inputs to multimodal models alongside text and images.

The research has evolved from supervised learning across narrow tasks to self-supervised methods that can scale. These architectures leverage both unlabeled brain data and pretrained semantic models, thereby resolving the generalization and data efficiency issues previously described.

1.3 Dissertation Contributions and Structure

This dissertation addresses the limitations of task-specific EEG analysis and makes the following contributions:

Architectural Innovations for General EEG Representations

It introduces novel neural architectures that learn generalizable representations from raw EEG signals:

- A frequency-aware multi-band encoder that processes delta, theta, alpha, and beta bands independently through parallel convolutional encoders, followed by residual CNN fusion for motor imagery classification. This architecture achieves improvements over handcrafted features and baseline deep learning models.
- Subject-conditional representation learning: A learnable subject layer that explicitly models inter-subject variability, enabling the network to capture subject-specific neural dynamics while

maintaining a shared representational space across individuals.

- A multi-layer transformer architecture adapted for variable-length EEG sequences that incorporates learnable positional embeddings and bidirectional GRU-based temporal feature extraction to handle the non-stationary nature of neural signals.

- Vector-quantized neural tokenization: A discrete representation learning framework that converts continuous EEG segments into reusable brain tokens via vector quantization with spectral reconstruction objectives. This enables compact downstream models and modular training paradigms.

Semantic Decoding from Neural Signals

It presents one of the first end-to-end systems for open-vocabulary EEG-to-Text decoding:

- Cross-modal alignment architecture: A two-stage training approach that aligns learned EEG representations with pre-trained language model embeddings (BART), enabling semantic decoding without vocabulary constraints.

- GPT-4 refinement pipeline: Integration of large language models for post-processing, improving grammatical coherence and semantic accuracy of decoded sentences.

- State-of-the-art performance: Our architecture achieves BLEU-1 of 42.75%, ROUGE-1-F of 33.28%, and BERTScore-F of 53.86% on the ZuCo datasets, representing improvements over previous state-of-the-art methods.

- Validation across 30 subjects: Comprehensive evaluation demonstrating consistent performance across diverse individuals engaged in natural reading tasks, with analysis of inter-subject variability and representational clustering.

Evaluation Methodologies and Empirical Analysis

It establishes evaluation frameworks and provide empirical insights:

- Multi-paradigm benchmarking: Comprehensive evaluation across three distinct BCI paradigms—motor imagery (4-class classification), emotion recognition (valence/arousal), and semantic decoding—demonstrating the versatility of learned representations.
- Quantification of generalization gaps: Systematic analysis revealing that traditional subject-dependent models suffer 19% accuracy degradation in cross-subject scenarios, with detailed characterization of session-to-session variability through PCA visualization and statistical analysis.
- Semantic-aware evaluation metrics: Introduction of BERTScore to EEG decoding evaluation, providing sentence-level semantic similarity assessment that better aligns with human perception compared to purely lexical metrics (BLEU, ROUGE).

Practical Deployability Advances

It addresses real-world deployment constraints for BCI systems:

- Consumer-grade hardware validation: All methods evaluated on dry-electrode EEG devices rather than clinical-grade wet electrodes, demonstrating robustness to noise and practical usability for everyday applications.
- Robotic control: Motor imagery-based control of robotic systems, demonstrating end-to-end pipeline from EEG acquisition through classification to robotic actuation, validating practical feasibility of BCIs for assistive technologies and human-robot interaction.

- Discrete brain tokens: Demonstration that vector-quantized representations enable lightweight downstream models with frozen encoders and small task-specific heads, suitable for resource-constrained devices and real-time applications.

Together, these contributions demonstrate that representation learning can transform EEG-based BCIs from task-specific pipelines into generalizable, interpretable, and deployable systems that bridge neural activity with human-interpretable semantics.

1.3.1 Structure

This dissertation is organized into five parts that progressively build from understanding current limitations to proposing generalizable solutions.

Part I: Theoretical Foundations establishes the theoretical and practical context for this work. Chapter 1 motivates the need for representation learning in EEG-based BCIs and introduces the challenge of bridging neural activity with semantic meaning. Chapter 2 surveys the fundamentals of BCIs, EEG signal processing, representation learning theory, and neural signal decoding in the era of foundation models. This chapter provides the theoretical and empirical foundations necessary for understanding the technical contributions in subsequent chapters.

Part II: From Task-Specific to General Representations examines classical BCI systems and their fundamental limitations. Chapter 3 implements traditional approaches to motor imagery classification and emotion recognition using handcrafted features and machine learning classifiers. These experiments establish baseline performance and reveal critical limitations. Models require per-subject training, fail to transfer across tasks, and exhibit severe performance degradation across subjects and sessions. Chapter 4 systematically analyzes these limitations, including the 19% drop in cross-subject accuracy and session-to-session variability, providing empirical motivation for learning invariant, transferable representations.

Part III: Learning Continuous Representations introduces deep learning architectures that automatically discover features from raw EEG signals. Chapter 5 presents frequency-aware, multi-band encoders with residual networks for spatiotemporal feature learning and demonstrates that learned representations outperform handcrafted features in motor imagery tasks. Chapter 6 introduces an end-to-end architecture for open-vocabulary EEG-to-text decoding that aligns learned neural representations with pretrained language models. Using subject-conditional transformers and cross-modal alignment with BART, the system achieves state-of-the-art semantic decoding performance across 30 subjects. This demonstrates that meaningful linguistic content can be extracted from non-invasive brain signals during natural reading.

Part IV: Discrete Representations & Foundation Models addresses the integration of neural signals with modern foundation models through discretization. Chapter 7 introduces vector quantization for EEG, converting continuous signals into discrete brain tokens. These tokens enable compact codebook representations, downstream task transfer, and modular training with frozen encoders.

Part V: Conclusions summarizes the findings in Chapter 8, which examines trade-offs between continuous and discrete representations. It also discusses practical deployment considerations and and outlines future directions toward large-scale brain foundation models and multimodal brain-language architectures capable of treating neural activity as a first-class modality alongside vision and language.

Part VI: Related Contributions presents supporting studies conducted in parallel that established foundational principles essential to this thesis. These works include clustering, sensor-based teleoperation, biometric authentication, physics-constrained generation, and dataset integration. They contribute critical insights on representation learning, subject variability, evaluation methodologies, and cross-dataset generalization, which directly inform the architectures and approaches developed in this thesis.

If I have seen further it is by standing on the shoulders of giants.

Isaac Newton

2

Theoretical Foundations and Literature Review

2.1 Fundamentals of Brain–Computer Interfaces and EEG Signal Processing

2.1.1 Brain-Computer Interfaces

The concept of brain-computer interfaces emerged in the 1970s, with early research focusing on developing communication systems for individuals with severe motor impairments Nicolas-Alonso & Gomez-Gil (2012). The first BCIs relied on invasive techniques, including the implantation of electrodes directly into the brain. While these systems demonstrated the feasibility of decoding neural activity to control external devices, their invasive nature posed a significant barrier to widespread adoption.

The advent of non-invasive techniques Cincotti et al. (2008), particularly those based on electroencephalography, marked a significant milestone in the field of brain-computer interfaces. Non-invasive BCIs utilize surface electrodes to record brain signals, eliminating the risks associated with surgical procedures. In recent decades, research in non-invasive BCIs has expanded rapidly, driven by advancements in signal processing, machine learning, and computational neuroscience. Today, BCIs are being employed in diverse fields, including assistive technologies, neurorehabilitation, gaming, and cognitive research, enabling users to control prosthetics, communicate via spelling devices, and interact with virtual environments Cecotti (2011); Zhuang et al. (2020).

Despite these advancements, several challenges remain, particularly in achieving reliable and robust performance in real-world scenarios. Figure 2.1 illustrates the spatial and temporal resolution of BCI measurement technologies. Invasive approaches, such as electrocorticography (ECoG) and microelectrode arrays (MEA), offer superior spatial resolution and millisecond-level temporal precision. In contrast, non-invasive techniques like EEG provide high temporal resolution but lower spatial fidelity due to the distance between the electrodes and the brain, leading to a more generalized representation

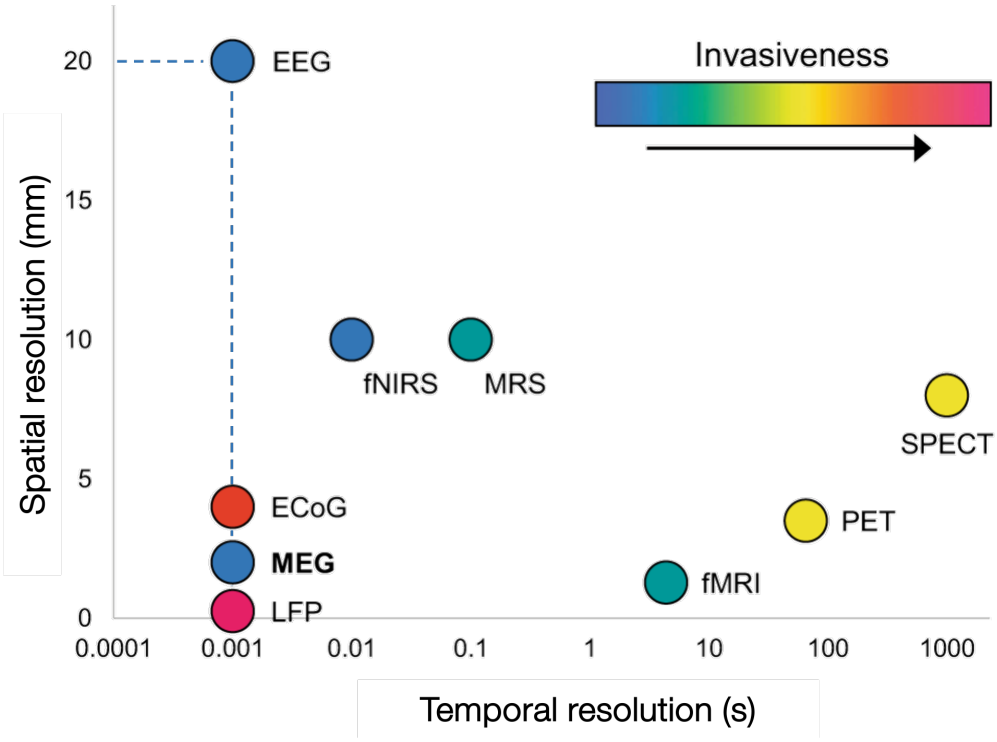


Figure 2.1: Spatial and temporal resolution by BCI technology Simmatis et al. (2023). Non-invasive methods are in blue: electroencephalography (EEG), magnetoencephalography (MEG), near-infrared spectroscopy (NIRS), functional magnetic resonance imaging (fMRI); invasive methods are in red: electrocorticography (ECoG), local field potential (LFP) recordings, micro-electrode array (MEA) recordings, and microelectrode (ME) recordings.

of cortical activity.

2.1.2 Electroencephalography

EEG is a non-invasive technique for recording brain electrical activity Binnie & Prior (1994). Signals are acquired by placing electrodes on the scalp, which detect voltage fluctuations caused by ionic currents within neurons. This time-series data provides valuable insights into brain function, supporting the study of various cognitive and physiological processes.

The acquisition process consists of several key steps: electrode placement, signal amplification, filtering, and digitization. Electrodes are typically positioned according to the International 10-20 sys-

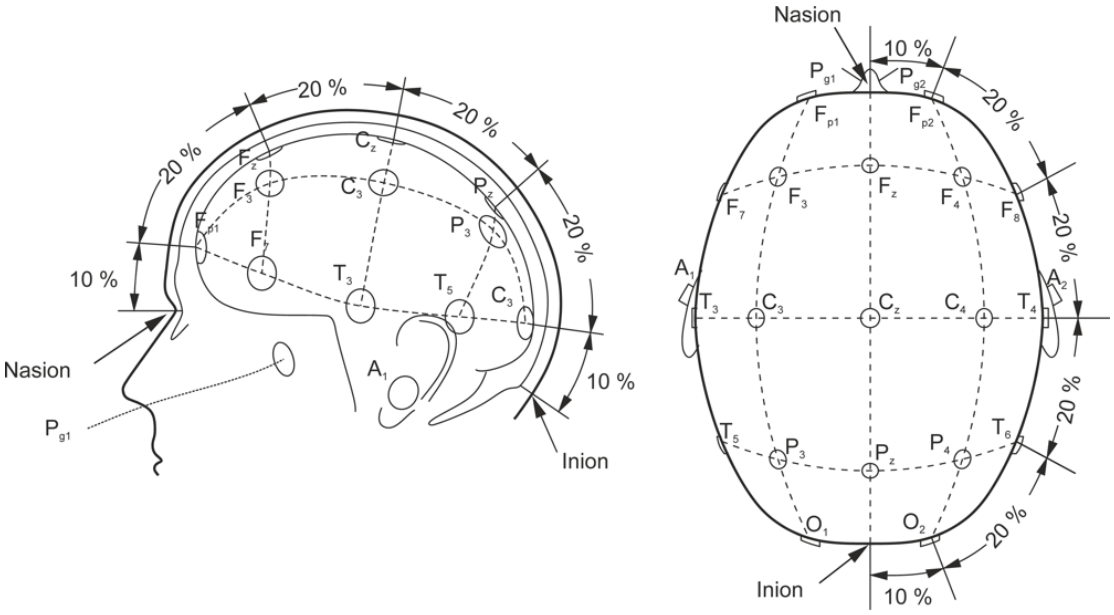


Figure 2.2: Position of electrodes with the International 10-20 system Nicolas-Alonso & Gomez-Gil (2012).

tem Herwig et al. (2003) (Figure 2.2), a standardized method ensuring consistency across studies.

EEG signals are characterized by high temporal resolution, typically in the millisecond range, making them well-suited for real-time applications. However, they are susceptible to noise and artifact sources from muscle activity, eye movements, and electrical interference (from the surrounding environment). Effective preprocessing and filtering techniques are crucial for removing artifacts Peng (2019); Sen et al. (2023).

Apart from implanted electrodes, which are beyond the scope of this discussion, EEG electrodes are classified into two main categories: wet and dry Mathewson et al. (2017). Wet electrodes use a conductive gel to minimize impedance at the electrode-skin interface, ensuring high-quality signal acquisition. However, applying the gel is time-consuming and may cause discomfort or skin irritation, making these electrodes less practical for long-term or daily use.

In contrast, dry electrodes eliminate the need for conductive gel, offering greater convenience and a faster setup, which makes them ideal for wearable and portable EEG systems. However, they generally

have higher impedance and are more susceptible to motion artifacts and signal degradation Lopez-Gordo et al. (2014).

EEG-based BCIs face several key challenges that must be addressed to enhance their efficacy and usability.

One major issue is the susceptibility of EEG signals to noise and artifacts, which can arise from muscle movements, eye blinks, and electrical interference. Developing robust preprocessing methods capable of filtering out noise while preserving relevant neural signals remains essential Scheer et al. (2005).

Another challenge stems from the significant inter-individual variability in EEG signals, influenced by differences in brain anatomy, physiology, and neural activity patterns. This variability complicates the development of generalized models that perform consistently across users Saha & Baumert (2020).

Additionally, BCIs must adapt to changes in neural activity over time, whether due to learning effects, fatigue, or other factors. In this context, incremental learning techniques that enable continuous model updates based on new data play a crucial role Giraud-Carrier (2000).

Finally, for BCIs to be practical, they must be comfortable for prolonged use and easy to set up. Advances in dry electrode technology and user-friendly interface designs are therefore fundamental to improving their usability Hairston et al. (2014).

2.1.3 EEG Signal Decoding Pipeline

The analysis of EEG signals follows a structured pipeline comprising several steps to ensure data reliability and interpretability. The process begins with **acquisition**, which encompasses the setup of the EEG device and the definition of the acquisition protocol. This is followed by **preprocessing** step that aims at eliminating noise and artifacts through various signal enhancement techniques. Subsequently, **feature extraction** is performed to derive relevant characteristics from the EEG signal. The

specific methods employed in this step depend on the application and may include time-domain analysis, frequency-domain analysis, or advanced nonlinear transformations. Finally, the **classification** step applies machine learning algorithms or other computational techniques to identify meaningful patterns within the processed data.

The pipeline is sketched in Figure 2.3, and the following subsections provide a detailed description of each step.

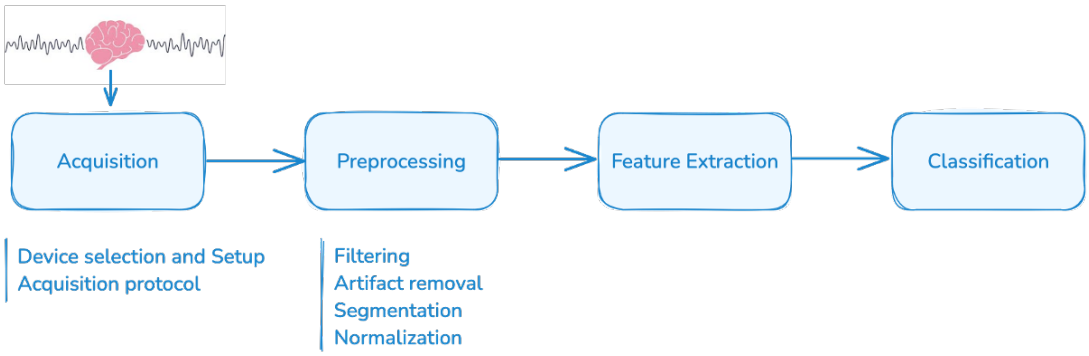


Figure 2.3: Overview of the EEG signal processing pipeline, from raw data acquisition through preprocessing and feature extraction to classification.

Acquisition

This step encompasses the choice and the setup and calibration of the EEG devices, as well as the implementation of a structured acquisition protocol designed to optimize signal quality and ensure consistency in data collection.

Devices selection and Setup. The choice of EEG devices and their setup are fundamental for capturing brain activity, as they directly impact the quality and reliability of the recorded signals. These devices consist of multiple electrodes placed on the scalp according to standardized systems, such as the International 10-20 system, ensuring consistent positioning across studies. The electrodes detect

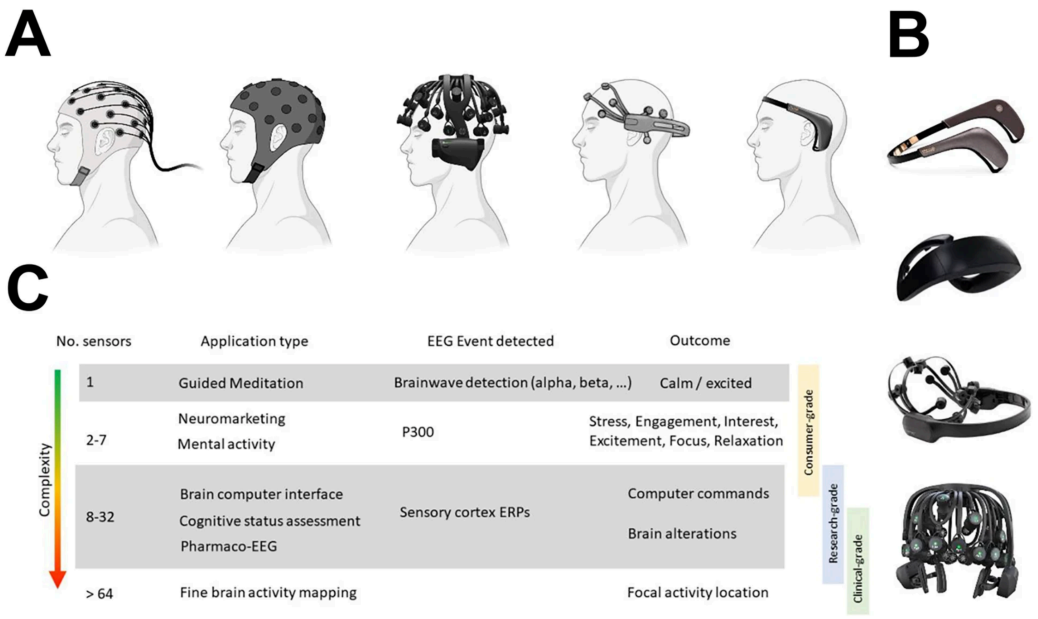


Figure 2.4: EEG data acquisition hardware varies widely in design and application Sugden et al. (2023). Panel (A) illustrates a range of setups from a traditional medical EEG system with numerous wired electrodes to various research and consumer wearables. Panel (B) shows some EEG wearables, including the Muse 2, Neurosity Crown, EPOC X, and Quick-32r. Panel (C) provides an overview of these devices, detailing their sensor counts, common applications, EEG characteristics, and associated results.

voltage fluctuations caused by ionic currents in the brain, which are then amplified, filtered, and digitized for further analysis.

EEG devices can vary in the number of electrodes they use. High-density systems, with 64, 128, or even more electrodes, provide superior spatial resolution and are typically employed in research settings Ferree et al. (2001). On the other hand, low-density systems, which use fewer electrodes, are commonly found in clinical and consumer applications due to their simplicity and ease of use. When selecting an EEG device for a particular application, it's essential to balance spatial resolution with usability

Data transmission methods are another important aspect of EEG devices. Many modern EEG systems use wireless technologies, such as Bluetooth or Wi-Fi, to transmit data to a paired computer

or mobile device for real-time processing and analysis. This wireless capability enhances user mobility and comfort, making EEG a valuable tool for everyday applications.

Figure 2.4 illustrates EEG data acquisition systems, ranging from traditional medical setups to research and consumer wearables, along with their specifications and applications.

Acquisition protocol. EEG data acquisition protocols are essential for obtaining high-quality recordings that enable accurate analysis and interpretation of brain activity. These protocols involve several key steps, each aimed at optimizing the reliability of the collected data.

Subject preparation is the first step, where participants are prepared for the recording process. They are seated comfortably in a controlled environment, with movement minimized to reduce artifacts that could interfere with the data collection. Clear instructions are given to help participants relax and remain still during the session, as muscle contractions and other movements could introduce unwanted noise into the EEG signals, thus compromising data integrity Moyer et al. (2017).

Following subject preparation, the electrode placement step is performed, during which electrodes are carefully placed on the scalp using standardized systems, such as the International 10-20 system. This structured method ensures consistency in electrode positioning across sessions and participants, which is crucial for maintaining reproducibility in research Yadav & Maini (2023). Proper placement enhances the quality of the recorded signals by optimizing spatial resolution.

Next, a baseline recording is conducted, where signals are recorded to establish a baseline for the subject. This is done before the main recording session begins. During this phase, participants are asked to remain still, with their eyes either closed or open, providing a reference point for normal brain activity without task-related influences. This baseline is essential for distinguishing spontaneous brain activity from responses triggered by specific tasks.

Following the baseline, the task instruction provision step ensures that participants receive clear instructions tailored to the study's objectives. The guidelines may vary, ranging from tasks involving visual or auditory stimuli to motor imagery or cognitive challenges. Clear instructions are crucial to

ensure participants effectively engage with the tasks.

Finally, in the stimuli presentation step, stimuli are delivered through visual, auditory, or tactile means. Specialized software manages the delivery of stimuli, ensuring precise timing and synchronization with the EEG recordings. This synchronization is vital for accurately correlating brain activity with specific stimuli or tasks Peksa & Mamchur (2023).

During the data recording phase, EEG signals are continuously recorded throughout the task or stimulus presentation. The duration of these sessions can range from minutes to hours, depending on the study design. To maintain participant engagement and data quality, short breaks are included to prevent fatigue Chaudhary et al. (2016). To facilitate subsequent analysis, event markers or triggers are used to annotate the EEG recordings with time stamps corresponding to specific stimuli or task events. These markers are invaluable for extracting event-related potentials (ERPs) and other task-specific features from the recorded data.

Finally, after the recording session is completed, post-session procedures involve debriefing participants and addressing any questions they may have about their experience. The recorded EEG data are then securely stored for preprocessing and analysis, marking the transition from raw data collection to detailed examination and interpretation.

Preprocessing

Preprocessing plays a fundamental role in EEG data analysis by enhancing the reliability of recorded signals through noise and artifact reduction. By transforming raw EEG data into a cleaner and more interpretable form, preprocessing is essential for accurate feature extraction and subsequent analysis. Techniques such as filtering, artifact removal, segmentation, and normalization significantly improve signal quality, ultimately enhancing the performance of EEG-based brain-computer interfaces.

The preprocessing pipeline typically starts with filtering, where the raw EEG signals are band-pass filtered between 0.5 Hz and 48.5 Hz using a 4th-order Butterworth filter. This step effectively ad-

dresses two key issues: low-frequency drifts caused by baseline shifts and slow physiological processes, and high-frequency noise, which may include power line interference and muscle artifacts. The careful selection of filter parameters is crucial, as it ensures that relevant neural information is preserved while noise frequencies are reduced.

Next, artifact removal is applied to address non-neural activity that often contaminates EEG signals, such as muscle contractions, eye blinks, and other physiological noise. Independent Component Analysis (ICA) is widely employed to decompose the EEG data into independent components Makeig et al. (1995). This method enables the identification of components corresponding to artifacts, which can then be selectively removed or corrected, thereby improving the quality of the remaining signal. The success of this step depends on accurately identifying artifact-related components, ensuring that meaningful neural signals are preserved.

Following artifact removal, segmentation is performed on the preprocessed EEG signals, dividing them into temporally overlapping windows. A common window length is two seconds, with a one-second overlap, although this may vary depending on the specific analysis requirements. Segmentation is crucial for preserving the temporal dynamics of brain activity, which is essential in many applications such as emotion recognition and motor imagery, where the timing of neural activity is critical for accurate interpretation. This step also facilitates feature extraction by dividing the continuous signal into manageable chunks for analysis.

Finally, normalization is applied to each segment of the EEG signal to account for variability introduced by factors such as electrode placement, scalp impedance, and individual neurophysiological characteristics. By subtracting the mean and dividing by the standard deviation of the signal, normalization ensures that all EEG channels are on a comparable scale. This uniformity is important for machine learning models, which rely on consistent input data across multiple electrodes. Additionally, normalization minimizes the impact of amplitude variations caused by non-neural factors, thus enhancing the reliability of the extracted features.

Feature Extraction

Feature extraction in EEG signal processing involves transforming raw EEG data into a set of meaningful features that can be used for classification or regression tasks. The importance of feature extraction lies in its ability to represent the brain's activity accurately while minimizing the influence of noise and artifacts. The choice of features significantly impacts the overall performance of BCIs, as it determines how well the extracted features reflect the brain's state during the task, whether it is motor imagery, emotion recognition, or cognitive state monitoring. For instance, in motor imagery classification, Common Spatial Pattern (CSP) Lu et al. (2010) is often used to enhance class discrimination by projecting EEG signals onto spatial filters that maximize variance between different motor imagery classes. In contrast, for emotion recognition or cognitive state monitoring, frequency-domain features like power spectral density (PSD) in various frequency bands (e.g., delta, theta, alpha, beta, and gamma) are commonly employed Wang et al. (2015), as they capture key neural oscillations associated with different cognitive and emotional states.

Broadly, the extracted features can be categorized into three main types: time-domain features, time-frequency domain features, and spatial features Hu & Zhang (2019); Al-Fahoum & Al-Fraihat (2014).

Time-domain features are directly derived from the EEG signal and capture various statistical properties that describe the signal's behavior over time. Commonly employed time-domain features include the mean and variance, which respectively represent the central tendency and the variability of the signal. The standard deviation is another important metric, providing insight into the dispersion of the signal values from the mean. Higher-order statistical measures, such as skewness and kurtosis, are also frequently used to characterize the shape of the signal's distribution. Skewness indicates the presence of asymmetry, while kurtosis measures the peakedness of the distribution, both of which are valuable for detecting non-normal or transient neural events that may carry relevant information.

Spatial features, on the other hand, are derived from the spatial distribution of EEG signals across different electrode locations on the scalp. These features capture the brain's spatial activity patterns, which can be used to distinguish between different cognitive or motor states. One of the most widely used spatial feature extraction methods is Common Spatial Patterns (CSP). CSP is particularly effective for tasks such as motor imagery classification because it maximizes the variance between different classes by projecting the signal onto spatial filters that emphasize class-relevant activity. CSP works by identifying linear projections of multichannel EEG data that optimize the discrimination between two conditions, making it a powerful tool for binary classification problems in BCIs. However, its application can be extended to multi-class problems using variations of the algorithm. The effectiveness of CSP in improving signal separability has made it a cornerstone method for spatial feature extraction, particularly in motor imagery and motion-related BCI applications.

Classification

The classification step is a fundamental step in EEG signal processing, where the features extracted from the raw EEG data are used to assign the data to specific categories or brain states. This step typically involves the application of various machine learning models to classify the EEG signals based on the extracted features. The choice of classifier depends on the complexity of the task, the size and dimensionality of the data, and the performance requirements of the BCI system.

In addition to selecting the appropriate classification approach, evaluating the performance of the classifier is essential. This is where the use of evaluation metrics becomes critical. These metrics offer a quantitative measure of how well the classifier is performing and provide insights into its accuracy and reliability. Moreover, classification metrics are often employed to give a more nuanced analysis of the model's effectiveness, allowing for a deeper understanding of its ability to differentiate between classes and how it manages false positives and false negatives. In the case of regression tasks, additional metrics (regression metrics) are used to assess how well the model predicts continuous values, providing

further insight into its performance and accuracy.

2.1.4 Machine Learning in EEG

Machine learning represents an effective approach for decoding EEG signals in BCIs. Among traditional methods, linear classifiers such as Linear Discriminant Analysis (LDA) and Support Vector Machines (SVM) have been widely adopted due to their simplicity and effectiveness. These techniques rely on handcrafted features extracted from EEG signals Shoeibi et al. (2021), including power spectral density, common spatial patterns (CSP), and wavelet coefficients, which capture relevant information for classification tasks.

In recent years, modern approaches, particularly deep learning, have played an increasingly important role in EEG signal analysis Craik et al. (2019); Gao et al. (2021). Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Attention-based networks, have demonstrated remarkable effectiveness in extracting features directly from raw EEG data. By capturing spatio-temporal patterns in EEG signals, these methods have led to significant improvements in classification accuracy and robustness.

Deep learning models, such as CNNs, are particularly effective at learning spatial hierarchies in EEG data by applying convolutional filters across the signal. They have been successfully employed in various BCI applications, including motor imagery classification, emotion recognition, and mental workload estimation. CNNs automatically learn important patterns from raw EEG signals, reducing the need for complex, manual feature design, and streamlining the model development process Fahimi et al. (2019); Sakhavi et al. (2018).

Recurrent Neural Networks, along with their variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are effective at capturing temporal dependencies in sequential data. These networks retain a memory of previous inputs, making them well-suited for tasks that require an understanding of temporal dynamics, such as continuous emotion monitoring

and sequential decision-making in BCIs Tortora et al. (2020); Zhang et al. (2018).

Attention mechanisms, often integrated with RNNs or Transformers, have further enhanced the capabilities of deep learning models in EEG analysis. By focusing on the most relevant parts of the signal, attention-based models improve both interpretability and performance. These mechanisms are particularly beneficial for applications that demand fine-grained analysis, such as decoding complex mental states and assessing cognitive load Li et al. (2020); Noah et al. (2020).

2.2 Foundations of Representation Learning

Unlike traditional feature engineering, where domain experts manually design input features, representation learning algorithms learn useful feature transformations as part of the model training process. This change has been crucial to the development of deep learning since it allows models to identify and arrange important patterns in data—like text, images, or signals—without the use of specially created features. These learned features, or representations, often reside in an abstract latent space. A good representation preserves important explanatory factors of the input data in a compact, information-rich form that facilitates subsequent tasks. Generally, high-quality representations should encode as much relevant information as possible, be low-dimensional enough to filter out noise, and generalize well (e.g., be transferable to new tasks).

Early approaches to representation learning include linear dimensionality reduction and manifold learning techniques. Classical methods, such as principal component analysis (PCA) Abdi & Williams (2010) and linear discriminant analysis (LDA) Balakrishnama & Ganapathiraju (1998), aimed to project data into a lower-dimensional subspace while preserving either variance or class separability. In the 2000s, nonlinear manifold learning methods Izenman (2012) emerged to uncover low-dimensional structures in high-dimensional data. These techniques demonstrated the benefit of transforming raw data into more tractable forms, laying the groundwork for future developments. However, a break-

through occurred in the mid-2000s with the development of deep neural networks for representation learning. In 2006, Hinton et al. Hinton et al. (2006) introduced deep belief networks with a greedy, layer-wise pre-training strategy, demonstrating that multiple layers of nonlinear features could be learned efficiently. This milestone shifted the focus toward deep representations. Each successive neural network layer learns higher-level abstractions (e.g., edges, shapes, and objects in images) by building on representations from previous layers. The success of deep learning over the last decade is closely related to these advancements in representation learning. In fact, the ability of deep models to automatically extract hierarchical features from raw data (pixels, waveforms, etc.) has largely replaced manual feature extraction in many areas.

2.2.1 Learning Representations: Theory and Techniques

In representation learning, we aim to find data transformations that reveal its underlying structure. Formally, a learned representation can be thought of as a mapping from the raw input, $x \mapsto z$ such that the raw input, x, is transformed into a new feature vector, z, in latent space. An ideal z preserves the meaningful information of x in a more organized way, such as by separating factors of variation, which is useful for prediction and other tasks. Neural networks achieve this by learning layers of nonlinear functions that progressively refine the representation. Early layers capture local or low-level patterns, and deeper layers capture global or abstract concepts. An important theoretical principle behind deep representation learning is the manifold hypothesis, which proposes that high-dimensional data lie on a much lower-dimensional manifold; representations uncover the coordinates on this manifold, filtering out extraneous dimensions (noise).

Learning representations can be supervised, unsupervised, or self-supervised. In supervised feature learning, the model is guided by target labels to shape the representation. For example, a convolutional neural network might learn features optimal for classifying object categories. In unsupervised settings, no labels are provided. Techniques such as autoencoders Hinton et al. (2011) and principal

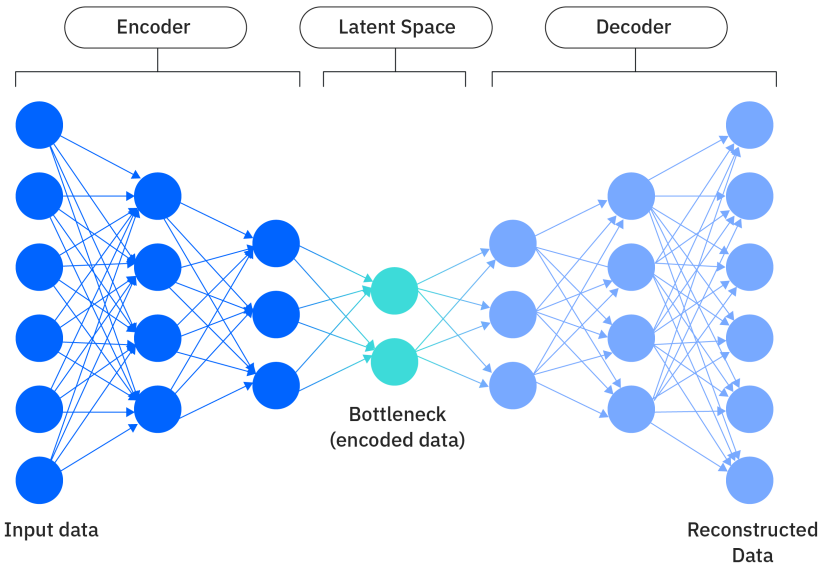


Figure 2.5: Architecture of an autoencoder. The encoder network compresses high-dimensional input data into a low-dimensional latent code (bottleneck), and the decoder network reconstructs the data from this code. By training to minimize reconstruction error, the autoencoder learns a compact representation capturing the data's important features.

components learn to represent the data based on its inherent structure (e.g., by reconstructing the input, as shown in Figure 2.5). Self-supervised learning is a modern approach to unsupervised learning. In this method, the system creates surrogate labels or pretext tasks from the data itself (e.g., predicting missing parts of an input or solving jigsaw puzzles made from images). The system then learns representations via supervised loss on these pseudo-labels. Self-supervised methods have notably enabled the use of vast amounts of unlabeled data to pretrain deep networks. This strategy has produced powerful, general-purpose representations in recent years. For example, in natural language processing and vision, models such as BERT Devlin et al. (2019) and SimCLR Chen et al. (2020) learn rich feature spaces by predicting masked words or distinguishing augmented image pairs instead of relying on human annotations.

Other representation learning frameworks include variational autoencoders (VAEs) Kingma &

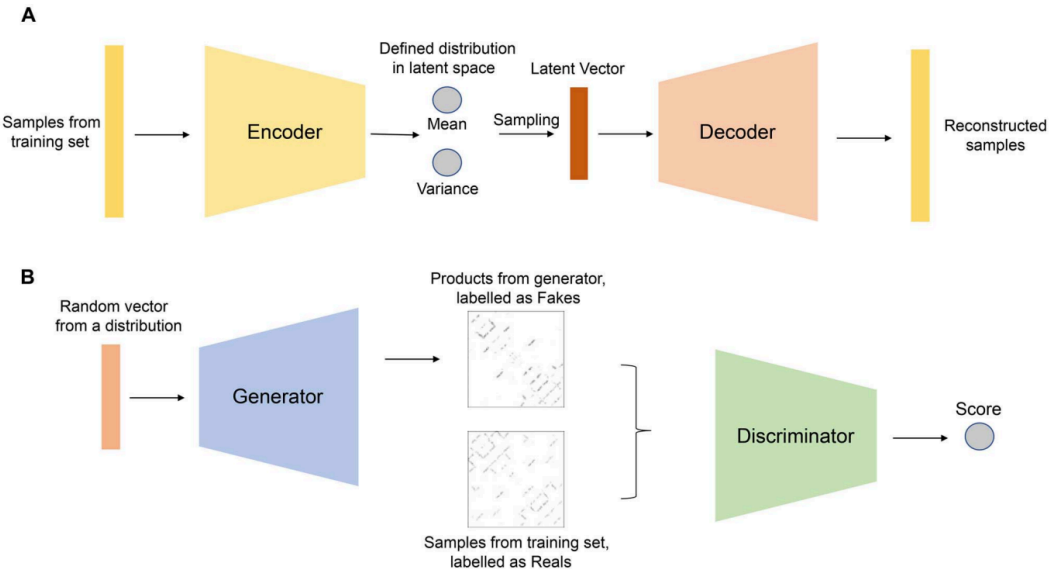


Figure 2.6: An illustration of the internal architectures of (A) VAEs and (B) GANs. Arrows represent corresponding data flow Ding et al. (2022).

Welling (2013) and generative adversarial networks (GANs) Goodfellow et al. (2020), showed in Figure 2.6. A VAE builds upon the traditional autoencoder by introducing a probabilistic latent space. Rather than mapping an input to a single code in a deterministic manner, the encoder learns a distribution — typically parameterized as a Gaussian with mean and variance vectors — over latent variables z. During training, a latent sample is drawn from this distribution and passed through the decoder to reconstruct the input. GANs, on the other hand, adopt a contrasting principle. Rather than directly reconstructing inputs, they train two networks: a generator that produces samples from random latent vectors and a discriminator that distinguishes between generated and real samples. Through this adversarial process, the generator learns to produce realistic outputs that match the underlying data distribution. In doing so, it implicitly learns a meaningful latent representation. Unlike VAEs, GANs typically prioritize sample realism over explicit latent structure, often resulting in sharper but less interpretable latent embeddings.

Another approach is metric learning, particularly contrastive Chen et al. (2020) and triplet-based

objectives Hoffer & Ailon (2015), which shape representations by enforcing proximity relationships between samples in latent space. In contrastive learning, for instance, the model learns to bring similar data points (positive pairs) closer together and dissimilar points (negative pairs) farther apart. This principle, implemented in frameworks such as SimCLR and MoCo, has become central to self-supervised learning in vision and speech recognition.

The common goal across all these frameworks is to transform raw, complex data into a latent representation space where relevant structures are geometrically and statistically defined. In this space, samples with similar underlying factors cluster together while those with different factors are separated. This organization improves predictive performance downstream and provides insight into the latent structure of the domain itself, revealing how visual, linguistic, or neural phenomena organize in continuous manifolds of meaning or dynamics, for instance.

2.2.2 Representation Learning for EEG Signals

In recent years, deep learning has increasingly been applied to EEG to automatically learn feature representations, mirroring the success seen in computer vision and natural language processing (NLP). One contributing factor has been the growing availability of larger EEG datasets and BCI benchmarks, which were historically difficult to obtain. With more data and computational power, since 2017, researchers have developed deep neural network models for EEG-based BCIs Hossain et al. (2023). These models learn layered representations of raw EEG data, which can improve performance in tasks such as motor imagery classification, seizure detection, and affective state recognition. Common architectures include convolutional neural networks (CNNs) adapted to multichannel time-series input, often with special spatial filters in the first layer to handle EEG electrode maps, as well as recurrent networks (RNNs) or transformers to capture temporal dependencies in EEG sequences. For instance, Schirrmeister et al. (2017) designed a deep convolutional neural network (ConvNet) for EEG that learned frequency-spatial filters in its initial layer, followed by deeper nonlinear feature extraction.

This yielded state-of-the-art BCI decoding accuracy at the time. Similarly, EEGNet Lawhern et al. (2018) introduced a compact CNN architecture specifically designed for EEG. It combines depthwise and temporal convolutions to learn frequency-specific features and has proven effective across different BCI paradigms. These deep networks essentially act as representation learners; the network's hidden layers form an embedding of the EEG signal that is discriminative for the task at hand (e.g., distinguishing between different imagined movements) and is often more robust than standard features.

A recent systematic review Guetschel et al. (2024) identified the wide variety of deep representation learning techniques used in BCI, highlighting the field's rapid growth. Notably, unsupervised and self-supervised learning approaches have emerged as a way to learn EEG representations without the need for large labeled datasets. Among the 81 BCI papers reviewed, autoencoders were the most prevalent unsupervised method: 34 studies employed some form of autoencoder to learn low-dimensional codes from EEG signals. These autoencoders can discover structure in EEG signals (e.g., denoising autoencoders can learn to remove artifacts), and their latent features can be used to train classifiers or facilitate transfer learning (adapting models across subjects or tasks). Beyond autoencoders, at least 13 studies have explored self-supervised learning (SSL) on EEG, with 10 of those appearing after 2021. This reflects how new the trend is: researchers are beginning to apply techniques such as contrastive learning, pretext task training, and Transformers (inspired by successes in vision and natural language processing) to brain signals. Recent works propose contrastive frameworks that learn EEG embeddings by grouping representations of the same trial under different augmentations or across subjects while separating representations of different classes or individuals. These SSL methods aim to learn invariant EEG features that generalize across variability (e.g., sessions or subjects), a crucial goal for BCIs since models often perform poorly when applied to a new user or slightly different context. Initial results are promising. Self-supervised EEG models have demonstrated improved robustness and cross-domain transfer (e.g., improving sleep stage classification by pretraining on unlabeled EEG

recordings). However, this line of work is very recent and still maturing.

2.3 Neural Signal Decoding in the Era of Foundation Models

Neural signal decoding has significantly evolved with the advent of foundation models, enhancing the understanding of brain activity in response to various stimuli. These models leverage large datasets and advanced machine learning techniques to improve decoding accuracy across diverse applications, from visual processing to BCIs. The following sections outline key advancements and implications of this evolution.

Recent progress in visual experience decoding illustrates the capabilities of foundation models in capturing complex neural representations. Studies using whole-brain fMRI analysis have shown that large pretrained models can reconstruct visual experiences with markedly higher semantic fidelity than traditional decoding methods, achieving improvements of over 40% in accuracy Wang et al. (2024b). These approaches integrate large-scale fMRI encoders with generative models, allowing a more comprehensive interpretation of visual stimuli that extends beyond the primary visual cortex to include associative regions such as the default mode network Wang et al. (2024b). This integration reveals that perception and imagination rely on distributed neural codes, a finding that underscores the power of representation learning in uncovering shared principles of brain organization.

Parallel developments have emerged in neuromodulation and symptom decoding, where foundation models trained on chronic electrophysiological recordings have demonstrated the ability to decode behavioral or clinical states without requiring patient-specific retraining Merk et al. (2025). Such models are capable of adapting to long-term fluctuations in neural activity and symptom expression, making them promising tools for closed-loop neuromodulation systems that adjust stimulation parameters in real time. Their capacity to generalize across individuals and recording conditions highlights a major step toward personalized yet scalable therapeutic interventions.

In the domain of language and cognition, multiple studies have shown that non-invasive brain recordings can capture multiple types of linguistic information Huth et al. (2016); Broderick et al. (2018); Caucheteux & King (2022), yet early decoding efforts were limited to reconstructing words or sentences within small, closed vocabularies Pereira et al. (2018); Dash et al. (2020); Moses et al. (2021), restricting their adaptability in open vocabulary contexts where linguistic diversity and semantic complexity are essential Wang & Ji (2022). Recent work shows that the brain encodes language into high-dimensional semantic spaces Gauthier & Ivanova (2018); Caucheteux & King (2022), similar to how pre-trained models such as BERT Devlin et al. (2018), BART Lewis et al. (2019), T5 Raffel et al. (2020), and GPT-4 OpenAI (2023) represent words as contextualized embeddings. Building on this parallel, several studies Wang & Ji (2022); Wang et al. (2023); Tang et al. (2023); Amrani et al. (2023b) have integrated brain decoding with language models to produce semantic brain embeddings, showing that linguistic representations can be inferred directly from neural activity Vaidya et al. (2022).

Recent advances have also focused on large-scale pre-training for neural signals. Building on self-supervised learning from vision and language, several studies have applied foundation-style pre-training to EEG and biosignals. Kostas et al. (2021) introduced BENDR, adapting contrastive learning to extract compact EEG features, while MMM Yi et al. (2023) and BIOT Yang et al. (2023) proposed topology-agnostic and tokenized pre-training frameworks enabling robust cross-subject learning. More recently, LaBraM Jiang et al. (2024) leveraged 2,500 hours of EEG recordings and a neural tokenizer to encode continuous signals into discrete latent codes for masked modeling, substantially improving transferability.

In parallel, multimodal large language models (MLLMs) have shown remarkable progress in integrating vision, text, and neural data. Models such as LLaVALiu et al. (2023), MiniGPT-4Zhu et al. (2023), and CogVLM Wang et al. (2024a) extend large language models with perceptual encoders to achieve cross-modal reasoning and alignment. These developments point toward a convergence between neural decoding and multimodal foundation modeling, paving the way for unified models

capable of jointly representing brain, language, and perception.

While the advancements in neural signal decoding through foundation models are promising, challenges such as individual variability and real-time processing constraints remain significant hurdles that need to be addressed for broader implementation in clinical and everyday settings Zhang (2025).

Part II

From Task-Specific to General Representations

Engineering is quite different from science. Scientists try to understand nature. Engineers try to make things that do not exist in nature.

Yuan-Cheng Fung

3

Classical EEG-Based Control Systems

3.1 EEG Acquisition and Motor Imagery Classification for Robotic Control

This section outlines the findings derived from assessing the feasibility of motor imagery BCIs in the control of robotic systems.

In particular, we examined the feasibility of employing portable EEG devices for extended duration in the context of robotic control scenarios. Our investigation is grounded in the hypothesis that portable EEG systems featuring dry electrodes, offer a practical and effective approach for the ongoing acquisition of EEG data, a pivotal requirement for real-world BCI applications.

The aim of our research was twofold: firstly, the development and testing of traditional learning-based techniques for classifying motor imagery tasks using signals acquired from a portable EEG device equipped with eight dry electrodes, and secondly, the practical experimentation of the generated models on a built miniature robotic vehicle simulating a wheelchair in which motor imagery tasks are matched with simple vehicle control commands.

Four motor imagery tasks have been considered: right hand, left hand, hands up, and feet lifting. Classifiers were designed to predict the identified motor activities. Additionally, the motor activities were grouped into two classes, with one class comprising the right hand and left hand, and the other containing the remaining two activities. This approach resulted in the generation of two-class classifiers.

The proposed approach involves segmenting the data acquired through the device, extracting Common Spatial Pattern (CSP) features, and subsequently training both a Support Vector Machine (SVM) and a K-Nearest Neighbors (KNN), one for each subject in the dataset. The dataset was constructed by involving 5 subjects ranging in age from 27 to 45 years. The collection of the EEG data has been made using a visual stimulation experimental setup.

Experimental results show an average accuracy of 36,32% for the multi-label classification, and

63,00% and 58,78% accuracy for the binary classification, respectively.

3.1.1 Material and Methods

3.1.2 Description of the MI Experiment for Robotic Control

The motor imagery experiment consists of four different motor imagery tasks and a resting state. Participants are instructed to vividly imagine four types of movements: moving the left hand, the right hand, both hands, and lifting the feet. These imagined movements have direct correlations with the robot control commands in the experiment: imagining moving the left hand corresponds to turning the left control, the right hand corresponds to turning the right control, both hands correspond to forward control, and lifting the feet corresponds to stop control. These tasks were selected based on their intuitive mapping to real-world robot controls, to enhance the usability and user adaptability of the system.

Five subjects (four males, and one female) aged between 27 and 45 years participated in this study. They declared not to have participated in any previous EEG experiment. Each participant was provided with a comprehensive overview of the study and gave their informed written consent. Participants were seated in a controlled environment, approximately 60cm away from a 27-inch LED display. To minimize external distractions and potential artifacts in EEG data, they were instructed to avoid unnecessary movements, particularly those involving the head and upper body, during the experiment. There are two parts to the experiment. First, subjects were asked to maintain a comfortable position with their eyes open and were instructed about the EEG device setup and the experimental procedure to perform. Participants were then asked to perform the four predefined MI tasks, with EEG data being continuously recorded. Each MI task involved the participant imagining the movement of the left hand, right hand, both hands, and lifting feet, corresponding to specific vehicle control commands. The tasks were presented as textual stimuli on the LED display for 5 seconds each. This

was followed by a 2-second blank screen and a 2-second fixation cross to allow participants to mentally prepare for the next task. The order of tasks was randomized for each participant to avoid order effects but kept consistent across participants for comparative analysis. Each session resulted in a 10-minute EEG recording covering the full range of tasks. The entire procedure was repeated twice for each participant, with a 5-minute break between sessions. This break was introduced to prevent fatigue and to maintain the quality of the motor imagery. The two different recordings denote training and testing data. For participant 1, the entire experiment was repeated two times on different days to test the repeatability of the EEG signals under similar conditions.

3.1.3 Data Acquisition

EEG data were collected using *abmedica*[1] Helmate device. Helmate is a non-invasive EEG Class IIA device certified according to the Medical Device Regulation (EU) 2017/745. The device comprises 10 dry electrodes, channels plus reference and bias, that follow the international 10-20 EEG system Klem (1999). The channels are located at FP1, FP2, Fz, Cz, C3, C4, O1, and O2, while the reference and bias electrodes are placed in the frontal region at AFz and FPz, respectively. Figure 3.1 reports the data signal acquisition details of abmedica Helmate device. Data is transmitted to a paired PC desktop via wireless Bluetooth at a sampling rate of 512 Sa/s, and raw data is recorded with the abmedica 'Helm8 Software Manager' software.

3.1.4 Data Pre-Processing

We preprocessed the entire MI-BCI dataset using the MNE-Python package Gramfort et al. (2013), which is openly available. The pre-processing phase, shown in Figure 3.2, consists of multiple steps:

[1] https://www.abmedica.it/

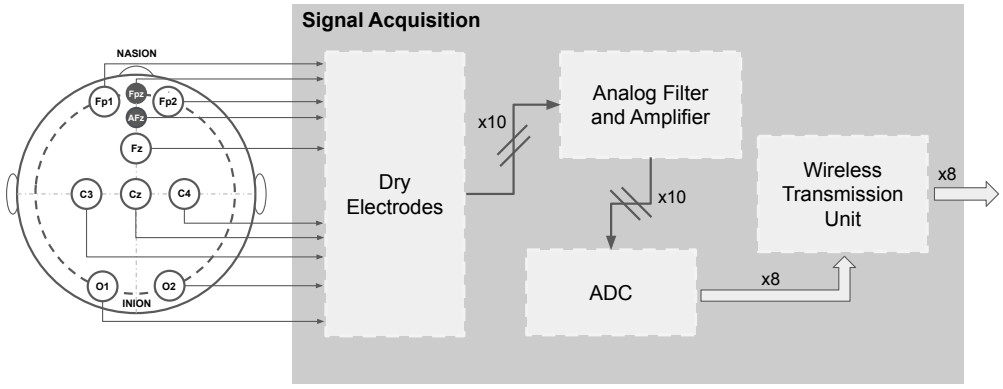


Figure 3.1: Acquisition process of EEG device Helmate8 from *abmedica*.

Filtering The EEG data is filtered between 0.5 and 48.5 Hz using a 4th-order band-pass Butterworth filter to mitigate the effects of noise, signal artifacts, and power-line interference.

Artifact Removal Independent Component Analysis (ICA) Comon (1994) is applied to the filtered signals to remove the common physiological artifacts, such as eye blinks, eye movements, and muscle contractions, more accentuated by dry electrodes and major freedom of movements.

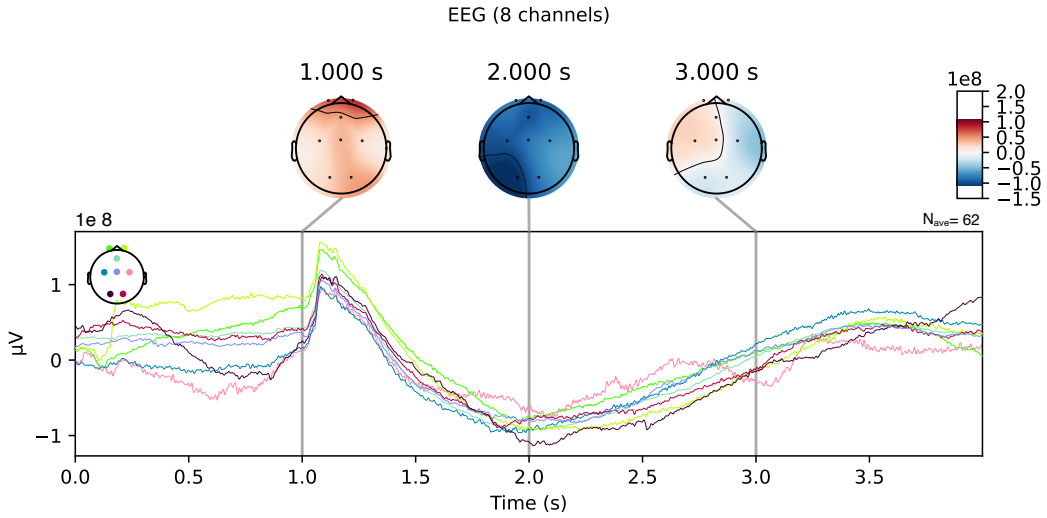
Windowing The preprocessed signals are segmented in windows of 2.0s with an overlap between subsequent segments of 0.5s. The preprocessed signals were further standardized on each electrode by subtracting the mean and dividing by the standard deviation, ensuring consistent scaling across all electrodes.

3.1.5 Feature Extraction and Classification

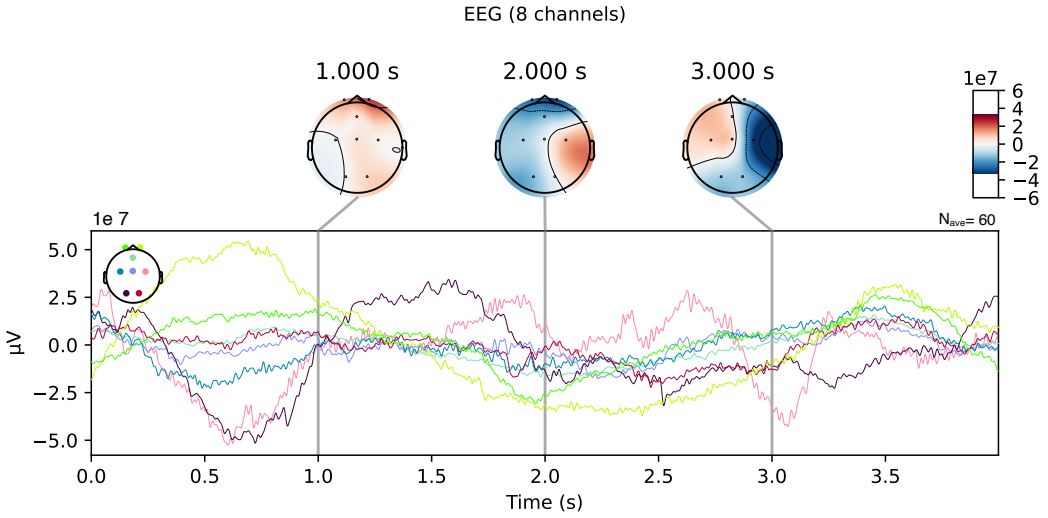
Each segmented EEG signal is filtered through a 12-band Filter Bank, with a 4 Hz bandwidth, equally spaced from 0.5 to 48.5 Hz. Then, they are fed to the Common Spatial Pattern (CSP) algorithm, obtaining a 96-feature vector representation for each window Ang et al. (2008). To classify these

EEG signals, we employed two machine learning algorithms: Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), thereby mitigating the computational burden that typically impacts Deep Learning solutions.

Both SVM and KNN were applied to classify the EEG data into four classes corresponding to the motor imagery tasks: left hand, right hand, both hands and lifting foot movement. Additionally, two binary classifications were performed to investigate the robustness of the classification models. First, both hands and lifting feet movements. Then, right-hand and left-hand movements. The training process was conducted individually for each subject, resulting in the creation of dedicated models for every participant.



(a) Raw signal.



(b) Signal after filtering and ICA artifact removal.

Figure 3.2: Proposed EEG pre-processing pipeline applied to the motor imagery dataset. The figure displays the signal at each processing stage: (a) raw EEG, (b) after band-pass filtering (0.5–48.5 Hz), (c) after ICA-based artifact removal, and (d) after windowing and standardization. For each stage, the averaged evoked response (left) and scalp topography at a representative time point (right) are shown across the 8 recording channels (FP1, FP2, Fz, Cz, C3, C4, O1, O2). The topographies illustrate the spatial distribution of activity, demonstrating progressive noise reduction while preserving task-relevant neural patterns.

Table 3.1: Macro average accuracy and standard deviation for each subject, across all the classes, for motor imagery classification. We report the results for multi-class and binary classifications.

	4 classes		2 classes (lift feet, both hands)		2 classes (right hand, left hand)	
Subject	**KNN**	**SVM**	**KNN**	**SVM**	**KNN**	**SVM**
1	40.00%	40.00%	65.56%	65.56%	67.78%	55.56%
2	38.95%	38.95%	61.11%	56.67%	52.22%	50.00%
3	34.74%	26.36%	60.00%	63.33%	58.89%	50.00%
4	33.78%	36.84%	56.67%	55.56%	57.78%	65.56%
5	35.79%	37.86%	78.89%	66.66%	72.22%	57.78%
avg ±std	36.65% ±2.41	36.00% ±4.93	64.45% ±7.76	61.56% ±4.58	61.78% ±7.22	55.78% ±5.77

3.1.6 Results and Discussion

3.1.7 Experimental Results

The performance of the models is evaluated in terms of the macro-average accuracy, which represents the average accuracy across all classes, and the corresponding standard deviation, which is calculated for each subject. We evaluate the classification performance for both multi-class (4 classes) and binary (2 classes) scenarios.

4-class classification

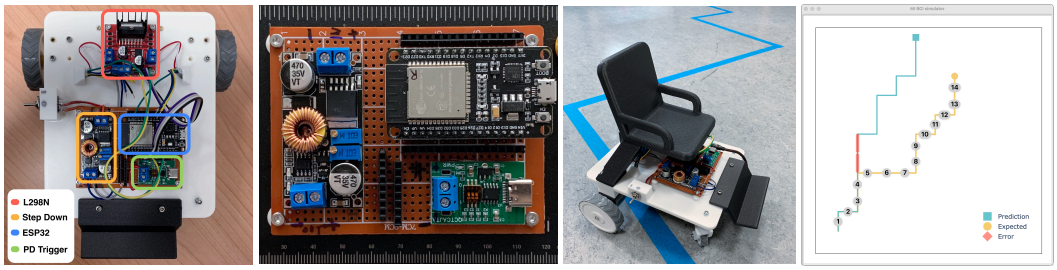
Table 3.1 summarizes the macro average accuracy and standard deviation across subjects for classifying MI tasks into four different classes. The results suggest that although the accuracy in distinguishing between all four classes is modest, it represents a promising first step in the development of MI-based BCI systems for vehicle control. The challenges of accurately classifying multiple motor imagery tasks are expected, given the subtle differences in EEG patterns associated with each task.

Binary classification

In addition to four-class classification, we also investigated performance in a binary classification setting. Two binary scenarios were considered: (1) "lift feet" and "both hands" tasks, and (2) "right hand" and "left hand" tasks. The results for the binary classification tasks are also shown in Table 3.2. Notably, the system achieved higher accuracy in the binary classification, with an average accuracy of approximately 64.45% for KNN and 61.56% for SVM when classifying the "lift feet" and "both hands" tasks. Similarly, the system achieved an average accuracy of about 61.78% for KNN and 55.78% for SVM in recognizing "right hand" from "left hand" tasks. This improvement in accuracy can be attributed to the reduced complexity of binary classification compared to multi-class classification.

3.1.8 Robot Vehicle Simulation

To simulate the results of our study, we used a scaled-down robotic vehicle model that integrates 3D printed components and key electronic elements (see Figure 3.3). This prototype should be understood in the context of our research scope - the controls implemented for this vehicle are relatively simple and do not reflect the complexities of real-world applications such as wheelchair control. The operation of the vehicle relies on a step-down DC-DC converter, an ESP32 microcontroller, and a USB-C power supply trigger, all carefully assembled on a prototyping board. Power is supplied by a portable battery that provides the smooth voltage regulation critical for consistent performance. Vehicle motion is provided by two DC-geared motors under the control of the L298N driver and the ESP32 microcontroller. This setup interprets signals via MQTT (Message Queue Telemetry Transport) and translates classifier predictions into corresponding motions. To better visualize the correctness of the vehicle path, we developed a GUI showing the correspondence between the planned path and the actual one (see Figure 3.3d).



(a) (b) (c) (d)

Figure 3.3: Scaled-down robotic vehicle employed for simulation: (a) support base with the various electronic components installed on it; (b) board with electronic components; (c) 3D printed components and key electronic elements; (d) GUI showing the correspondence between the planned path and the actual one.

3.1.9 Conclusions

In this work, we explore and validate the practical application of low-cost and minimally invasive EEG devices for the control of robotic systems. In the experiment, we collected EEG data from five participants and trained machine learning models on motor imagery tasks that can be used to control robotic systems, such as wheelchairs and robotic arms. The experimental results on the four-class task are not very accurate, while the results on the binary classification tasks are very encouraging with an average accuracy of about 61%, demonstrating the viability of using portable EEG devices with dry electrodes. Based on the results of our research, there is significant potential to expand the scope of BCI applications in the area of assistive technologies, such as wheelchairs driven by EEG-based systems and robotic arms controlled by neural signals, that provide new levels of independence for people with disabilities.

3.2 Emotion Personalization with Machine Learning using EEG Signals

Emotion recognition plays a vital role in various fields, including healthcare, affective computing, and human-computer interaction Ali et al. (2016); Bos et al. (2006); Alimardani & Hiraki (2020). One of the most relevant tasks in the field of emotion recognition is the measurement of valence and arousal dimensions, which represent the positive/negative and low/high activation levels of emotions, respectively. However, accurately and reliably classifying valence and arousal from brain signals poses significant challenges due to the complexity of neural responses and the inter-subject variability, which either refers to the differences in neural responses and patterns of brain activity related to emotions across different individuals Kim & Jo (2018). Each person has a unique brain structure, neural connectivity, and physiological characteristics, which can lead to variations in how emotions are represented and manifested in the brain Ekman et al. (1999). For example, genetic differences, life experiences, cultural backgrounds, and individual cognitive and emotional processes may influence the way emotions are experienced.

In this work, we address the challenge of personalizing BCI models utilizing a wireless consumer non-invasive electroencephalogram (EEG) device with dry electrodes. Our research investigates the effectiveness of three machine learning algorithms in classifying valence and arousal: k-Nearest Neighbors (k-NNs), Support Vector Machines (SVMs), and Artificial Neural Networks (ANNs). To achieve personalization, we adopt an incremental learning approach by progressively incorporating high-quality subject data during model training that is taken from the ground-truth. We compare the performance of the models before and after personalization. Results demonstrate that personalized models consistently outperformed the non-personalized models across all the classifiers we employed of a significant amount. Specifically, the integration of subject-specific data led to better discrimination of emotional states, resulting in accuracy improvements of 27.82% and 28.80% for valence and arousal, respectively, when compared to the non-personalized models.

In a nutshell, the main contributions of this work are the following.

- Addressing the personalization problem in non-invasive BCIs for valence and arousal classification using machine learning techniques.
- Experimentation on EEG data collected from a wearable 8-channel dry electrode device.
- Introducing an incremental learning framework that integrates subject-specific data into the training process.
- Highlighting the effectiveness and potential of personalization techniques in enhancing the accuracy of BCI models for emotion recognition.

3.2.1 Data: Valence and Arousal

Valence and arousal are two fundamental dimensions in the domain of emotional states. They represent key aspects of human subjective experiences and serve as important constructs in the field of emotion research.

Valence refers to the subjective quality or pleasantness of an emotional state, ranging from positive (e.g., happiness, joy) to negative (e.g., sadness, fear). It reflects the overall hedonic tone of an emotional experience, with positive valence indicating positive affect and negative valence indicating negative affect.

Arousal, on the other hand, pertains to the intensity or activation level of an emotional state. It captures the degree of physiological and psychological activation, ranging from low arousal (e.g., calm, relaxation) to high arousal (e.g., excitement, tension). Arousal reflects the energy, alertness, and readiness for action associated with an emotional experience.

Valence and arousal are often represented as a two-dimensional space known as the affective circumplex model Russell (1980). This model provides a framework for understanding and categorizing

Figure 3.4: Electroencephalogram device Helmate8 from *abmedica*. From left to right: outside and inside of the device, a dry electrode.

emotions based on their location in the valence-arousal space. Emotions can be positioned within this space according to their valence (positive or negative) and arousal level (low or high), allowing for a comprehensive characterization of emotional experiences.

The dataset used in the experimentation is a wearable MI-BCI-based Apicella et al. (2022) that has been acquired using an *abmedica* Helmate8 device, shown in Figure 3.4.

The dataset contains EEG recordings from a total of 26 participants. Each participant took part in a single session, during which an ad-hoc experimental protocol was implemented to elicit different emotional states Apicella et al. (2022). The experimental design was guided by the circumplex model of affect Russell (1980), which served as the theoretical framework. Specifically, the aim was to induce positive/negative valence and high/low arousal states by presenting a randomized sequence of International Affective Picture System (IAPS) images Lang et al. (2005). Participants were instructed to passively observe the images without actively attempting to regulate their emotions. Prior to the start of the experiment, detailed information regarding the study's purpose and methods was provided to the participants. They were also instructed to minimize their movements during the tasks, ensuring minimal interference with the EEG recordings. The EEG signals were acquired at a sampling rate of 512 Sa/s, transmitted and stored using Simulink (Matlab v. R2021b).

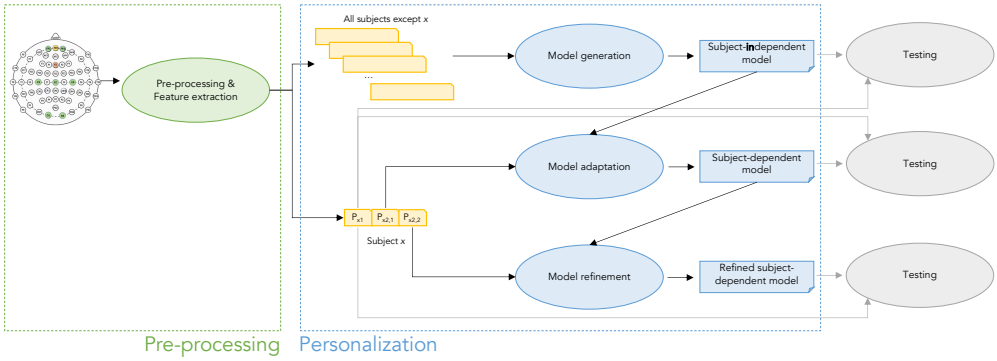


Figure 3.5: The proposed method to personalize valence and arousal classification models using EEG signals.

3.2.2 The Personalization Method

This work aims to investigate and address the personalization problem in BCIs for valence and arousal classification using machine learning techniques. The objective is to explore the effectiveness of incorporating subject-specific data and adapting machine learning models to individual users to enhance the accuracy and reliability of BCI systems for emotion recognition. To achieve this aim, the study focuses on the following two research questions:

- RQ1: Does personalization improve the accuracy and reliability of valence and arousal classification in non-invasive BCIs?
- RQ2: How do different machine learning algorithms compare in the context of personalized non-invasive BCI systems?

To address these research questions, this study aims to contribute to the understanding of the benefits and challenges of personalization in BCI systems, in the context of valence and arousal classification.

Figure 3.5 sketches the personalization method, which consists in two main phases.

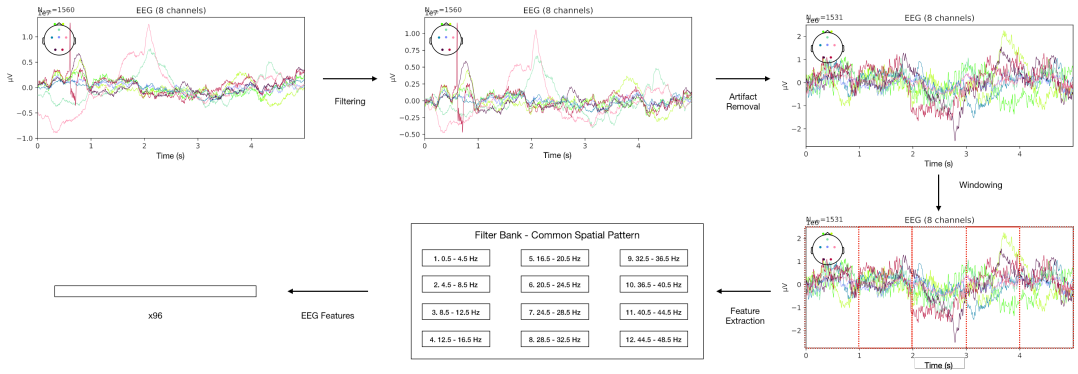


Figure 3.6: The proposed pre-processing and feature extraction method for EEG signals.

The Pre-processing phase aims at enhancing the quality of the EEG data by removing artifacts and extracting discriminative features.

The Personalization phase aims at tailoring a classifier for valence and arousal classification personalized to the target subject. The phase includes three steps. The first step derives a subject-independent model using the subjects in the dataset except for target subject *x*. The second step aims at adapting the subject-independent model to the subject *x*. The third and last step exploits incremental learning to facilitate the adaptation of the subject-dependent model to new, unseen, data from the subject *x*.

Pre-Processing Phase

The architecture of the proposed pre-processing and feature extraction approach is shown in Figure 3.6.

The pre-processing phase consists of 4 steps:

Filtering The EEG data is filtered between 0.5 and 48.5 Hz using a 4th-order band-pass Butterworth filter to mitigate the effects of noise, signal artifacts, and power-line interference. In addition, the selected range contains the traditional five EEG bands: delta (0-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (>30Hz).

Artifact Removal Independent Component Analysis (ICA) Comon (1994) is applied to the filtered signals to remove the common physiological artifacts, such as eye blinks, eye movements, and muscle contractions, more accentuated by dry electrodes and major freedom of movements.

Windowing The preprocessed signals are segmented in windows of 2.0s with an overlap between subsequent segments of 1s. For instance, 5 seconds of data, corresponding to 2560 samples, was split into four windows of 2.0s, composed of 1024 samples each. The total number of segments obtained was 6,124, among all the subjects.
The preprocessed signals were further standardized on each electrode by subtracting the mean and dividing by the standard deviation, ensuring consistent scaling across all electrodes.

Feature Extraction Each segmented EEG signal is filtered through a 12-bands Filter Bank, with a 4 Hz bandwidth, equally spaced from 0.5 to 48.5 Hz. The traditional five bands are divided into 12 sub-bands. In this way, each window is composed of 12 frequency bands, 8 channels, and 1024 samples. Then, they are fed to the Common Spatial Pattern (CSP) algorithm, obtaining a 96-feature vector representation for each window Ang et al. (2008).

Machine Learning Classifiers

To respond to RQ2, we selected three well-known machine learning algorithms: k-Nearest Neighbors (k-NN), support vector machines (SVM), and artificial neural networks (ANN).

k-Nearest Neighbor The k-NN algorithm is a simple yet effective non-parametric classification method. Given a new input sample, the k-NN algorithm finds the k-nearest samples in the training set based on a distance metric and assigns a class label to the new sample based on the majority class among its k-nearest neighbors.

The number of neighbors (k) was set to 1. The Euclidean distance metric was used to measure the similarity between instances.

Support Vector Machine (SVM) SVM is a powerful supervised learning algorithm used for classification tasks. SVM aims to find an optimal hyperplane or set of hyperplanes that separate the different classes with the maximum margin. SVM can handle linearly separable as well as non-linearly separable data by using the kernel trick. The kernel function transforms the input data into a higher-dimensional feature space, where the classes become separable.
A Radial Basis Function (RBF) kernel was selected, while the regularization parameter (C) was set to 1.

Artificial Neural Network (ANN) ANNs are computational models inspired by the structure and function of biological neural networks. They consist of interconnected nodes (neurons) organized in layers: input, hidden, and output layers. Each neuron receives input signals, performs computations, and passes the output to the next layer.
The ANN consisted of four fully connected layers [64, 64, 32, 2] with ReLU activation functions. It used the Adam optimizer with a learning rate of 1e-3, and the cross entropy loss function. The batch size was 32, and the total number of epochs was 200.

Personalization Phase

The objective of personalization is to build a model that is tailored to subject x.

The personalization phase generates three incremental models: from the most general one (subject-independent) to the one most tailored to subject x (refined subject-dependent model). The generated models are validated to determine their effectiveness. For this purpose, the samples belonging to subject x are divided into two equal partitions: P_{x1} and P_{x2}. P_{x1} serves as the testing set to evaluate the

performance of the models, while P_{x2} is used to adapt the model to subject x. The personalization phase involves the following three steps.

Model Generation The goal of this step is to establish a robust and generalized model that can serve as a starting point for subsequent personalization. The generated model, termed *subject-independent model*, is obtained using the data from all subjects except for the target subject x. Traditional machine learning classifiers may be employed to train the *subject-independent model*.

Model Adaptation This step focuses on subject x and aims to adapt the subject-independent model to the individual characteristics of the target subject x. The P_{x2} set is partitioned into $P_{x2,1}$ and $P_{x2,2}$, both of equal size. $P1_{x2}$ is utilized to fine-tune the *subject-independent model* generated in the previous step. By incorporating subject-specific data, the generated model, termed *subject-dependent model*, is tailored to capture the unique neural responses and patterns of brain activity associated with the valence and arousal dimensions.

Model Refinement This step aims to adapt the subject-dependent model more closely to the characteristics of subject x in order to improve its performance.

The incremental learning phase takes advantage of an additional subset of the subject x's data: $P_{x2,2}$. By incrementally integrating new data samples of the subject x into the model (termed refined subject-dependent model), it becomes more able at capturing the different subject's neural responses and enhancing its predictive accuracy for valence and arousal classification.

The validation procedure involved executing the 3 steps for all 29 subjects present in the dataset. For each subject, the procedure was replicated for the three machine learning algorithms described in Section 3.2.2 (k-NN, SVM, and ANN), resulting in a total of 9 models per subject.

Table 3.2: Macro average accuracy and standard deviation, across all the subjects, for valence and arousal classification. The best performance value is highlighted in bold.

	Valence			Arousal		
Classifier	Subject-independent	Subject-dependent	Refined subject-dependent	Subject-independent	Subject-dependent	Refined subject-dependent
k-NN	48.94% ±04.95	**65.36% ±04.32**	**76.76% ±03.83**	**49.06% ±07.19**	**68.97% ±09.35**	**77.86% ±07.88**
SVM	**49.81% ±05.05**	62.17% ±05.29	65.82% ±04.42	46.31% ±10.76	57.16% ±12.05	63.06% ±11.31
ANN	49.17% ±05.42	53.30% ±04.39	56.83% ±05.13	48.94% ±09.38	56.45% ±09.12	61.07% ±09.32
average	49.30 % ±05.14	60.27 % ±04.66	66.47 % ±4.46	48.10 % ±09.11	60.86 % ±10.17	67.33 % ±9.50

3.2.3 Experimental Results

All three models generated in each step are evaluated over the P_{x1} that serves as the testing set.

The performance of the models is evaluated in terms of macro average accuracy, which represents the average accuracy across all classes, and the corresponding standard deviation, calculated across all subjects.

Table 3.2 shows the average of the results obtained by all methods over the EEG dataset, for both valence and arousal. Table 3.3 shows the accuracy of both valence and arousal for the k-NN classifier (k-NN) for all the subjects.

The results demonstrate that the personalized models (i.e., subject-dependent and refined subject-dependent), on average, consistently outperformed the non-personalized models (i.e., subject-independent) in terms of macro accuracy.

The accuracy of the personalized models exhibited a significant improvement across all three classifiers, 17.17% for valence and 19.23% for arousal, indicating that the integration of subject-specific data led to better discrimination of emotional states.

Response to RQ1: Emotional states are more discriminable when the model used for their prediction is built using a personalized approach. For both valence and arousal dimensions, the k-NN classifier achieved the highest accuracy among the three classifiers in the personalized models. The subject-dependent and refined subject-dependent classifiers showed substantial improvements, achieving accuracy rates of 65.36% and 76.76% for valence, and 68.97% and 77.86% for arousal, respectively, with

Table 3.3: Macro average accuracy for each subject when k-NN is adopted during the proposed personalization procedure.

	Valence			Arousal		
Subject	Subject-independent	Subject-dependent	Incremented Subject-dependent	Subject-independent	Subject-dependent	Incremented Subject-dependent
001	47.50%	66.67%	78.33%	45.00%	71.67%	84.17%
002	51.67%	68.33%	86.67%	51.67%	61.67%	76.67%
003	49.17%	62.50%	77.50%	33.33%	74.17%	80.00%
004	56.67%	69.17%	76.67%	49.17%	65.73%	74.20%
005	50.00%	56.67%	76.67%	51.67%	60.00%	72.50%
006	47.50%	70.83%	76.67%	50.83%	62.50%	75.83%
007	55.83%	64.17%	79.17%	34.17%	46.87%	62.44%
008	50.00%	65.83%	75.83%	52.50%	75.83%	82.50%
009	50.83%	62.50%	80.00%	56.67%	75.00%	78.33%
010	50.00%	67.50%	75.83%	40.83%	83.33%	90.83%
011	55.00%	66.67%	79.17%	49.17%	70.00%	76.67%
012	53.33%	72.50%	74.17%	43.33%	70.00%	73.33%
013	51.67%	65.83%	80.00%	55.00%	60.00%	74.17%
014	50.83%	56.67%	72.50%	55.83%	72.50%	76.67%
015	45.83%	65.83%	75.00%	55.83%	68.33%	80.00%
016	46.67%	68.33%	75.00%	48.33%	55.83%	75.83%
017	45.83%	70.83%	80.83%	39.17%	80.00%	82.50%
018	53.33%	59.17%	65.00%	59.17%	60.00%	58.33%
019	54.17%	60.83%	75.00%	50.83%	78.33%	82.50%
020	45.00%	71.67%	78.33%	48.33%	62.50%	76.67%
021	35.83%	67.50%	78.33%	53.33%	80.83%	85.83%
022	43.33%	60.00%	77.50%	50.83%	81.67%	88.33%
026	40.00%	64.17%	80.83%	55.00%	83.33%	94.17%
027	50.00%	61.67%	71.67%	60.83%	72.50%	80.00%
028	40.83%	64.17%	75.00%	47.50%	60.83%	77.50%
029	51.61%	69.35%	74.19%	37.10%	59.68%	64.52%
average	48.94% ±04.95	65.36% ±04.32	76.76% ±03.83	49.06% ±07.19	68.97% ±09.35	77.86% ±07.88

improvements of 27.82% and 28.80% respectively.

Response to RQ2: When dealing with the prediction of emotional states, a kNN-based personalized model allows to obtain better results with respect to SVM and ANN.

3.2.4 Discussion and Limitations

The findings from our study highlight the ability to adapt BCI models to individual users, considering their unique neural responses and patterns of brain activity related to emotions. The inter-subject variability observed in emotional responses can pose challenges when attempting to develop generalized models. However, by personalizing the models, we can account for these individual differences and tailor the classification algorithms to better capture the specific emotional states of each subject. This individualized approach holds great promise for improving the overall performance and user experience of BCIs.

While our findings provide valuable insights into the benefits of personalization, further investigations with larger and more diverse datasets are necessary to validate the generalizability of our results. Expanding the dataset to encompass a broader range of demographic factors, cultural backgrounds, and emotional responses would enhance the robustness and applicability of the personalized models. Another area for future exploration is in the interpretation and explainability of personalized BCI models. As personalization techniques become more sophisticated, it becomes crucial to understand how these models arrive at their decisions. Developing transparent and interpretable models will not only increase user trust but also allow for better understanding of the underlying neural mechanisms associated with emotions.

Future research in emotion personalization with consumer non-invasive EEG devices, should focus on: 1) validating findings with larger and diverse datasets, 2) improving interpretability of personalized models, 3) integrating multimodal data for a comprehensive understanding of emotions, and 4) enabling long-term adaptation based on user feedback. These efforts will enhance the generalizability, and usability of personalized emotion recognition models, ultimately maximizing their potential in improving BCI performance and user experience.

In one word, to draw the rule from experience, one must generalize; this is a necessity that imposes itself on the most circumspect observer.

Henri Poincare

4

Limitations and Motivation for General Representations

The previous chapter talked about the old ways of using EEG to control systems. It showed how these systems have been used to identify different motor imagery and recognize emotions. While these models have shown good performance in some areas, they have also shown basic limitations that make it hard for brain-computer interfaces to be used more widely. This chapter looks closely at these limits and explains the theoretical and practical reasons for learning general-purpose neural representations.

4.1 Why Task-Specific EEG Models Fail to Generalize

The most common approach in processing neural signals has been to create features and train models for specific tasks. This method has produced functional BCIs, but there are major limitations when trying to expand these systems.

4.1.1 Non-transferability Across Tasks

There are different ways to process neural signals. Each of these solutions requires a complete redesign of the system and training from the beginning. Think about the examples in Chapter 3. The motor imagery and emotion recognition systems both use EEG signals, and they are processed in similar ways. For example, they both use bandpass filtering, artifact removal, and channel selection. However, the knowledge gained from each model doesn't directly apply to the other's task. A model that is trained to understand motor intentions cannot inform emotion classification, even though both tasks involve the same brain regions. This non-transferability makes it difficult to use in practice. Each new application requires different data to be collected, labeled, and prepared. As an example, researchers trying to expand from motor imagery to emotion recognition have to start from scratch. This inefficiency is particularly problematic in medical settings. There, different tests—like detecting seizures, tracking sleep, and assessing cognitive load—use the same neural signals. But each test needs its own model. This has led to many simple solutions that don't fully capture the complex information in

brain signals.

4.1.2 Subject Dependence and Overfitting

Neural signals exhibit substantial variability among individuals due to anatomical differences, variations in skull thickness, and unique neural dynamics. Our emotion personalization experiments clearly illustrate this limitation. Subject-dependent models achieve an average accuracy of 67.33% across three different architectures. However, this performance catastrophically degrades to 48.10% when evaluated in a subject-independent manner. This 19.23% drop in performance reveals that current approaches learn subject-specific artifacts rather than universal neural patterns. The models memorize unique features such as electrode impedance patterns, anatomical features, and cognitive strategies rather than extracting transferable neural representations. Consequently, deploying these systems to new users requires extensive calibration sessions, rendering them impractical for real-world applications where immediate functionality is essential. The potential of brain-computer interfaces as assistive technologies remains unfulfilled when hours of training data must be collected by users before the system can be used.

4.1.3 Session Variability

Even within the same subject, neural signals demonstrate significant non-stationarity across recording sessions. As demonstrated in the motor imagery experiments of Chapter 3, models trained on data from one session face significant challenges in maintaining performance when tested on subsequent recordings, even when experimental conditions remain nominally identical. This session variability is due to a number of factors. Electrode placement varies between sessions, even when cap positioning protocols are followed with care. It is important to note that impedance characteristics are subject to variation due to factors such as skin condition, temperature, and humidity. More fundamentally,

neural dynamics themselves evolve—factors such as fatigue, learning effects, and circadian rhythms all modulate brain activity patterns. It is important to note that users may develop different cognitive strategies over time, which can fundamentally alter the neural signatures of identical tasks.

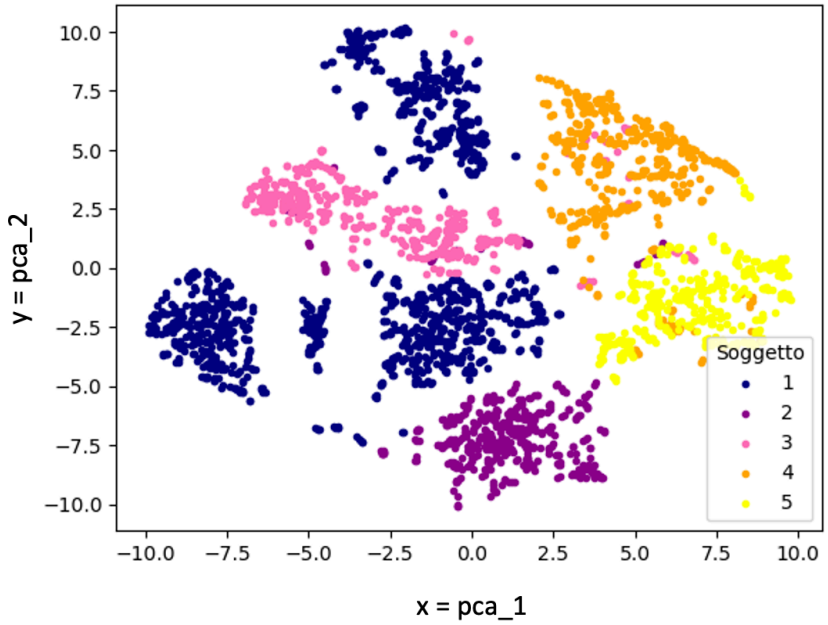


Figure 4.1: PCA projection of EEG raw signals across five subjects during motor imagery for robotic control. Each color corresponds to a different subject. For Subject 1, three distinct clusters are visible, corresponding to three recording sessions. Despite identical experimental conditions, the separation of these clusters demonstrates significant session variability, highlighting the non-stationarity of neural signals over time.

Figure 4.1 illustrates an example of this phenomenon. It visualizes the PCA projection of EEG raw signals collected from five subjects during motor imagery tasks for robotic control. For Subject 1, three separate recording sessions are represented by distinct clusters in the latent space. This reveals substantial variability in neural patterns across sessions, despite identical experimental conditions. This dispersion highlights how session-dependent variations can prevent generalization and reduce model reliability.

This temporal instability poses a significant challenge to the reliability of brain-computer interfaces. Calibrated systems can be unreliable, failing to meet performance expectations. This can lead

to frequent recalibration, which is both frustrating for users and limits practical deployment. Current task-specific approaches lack mechanisms to adapt to these temporal dynamics, treating each session as an independent problem rather than recognizing the underlying continuity of neural processes.

4.2 Toward Invariant and Transferable EEG Representations

The limitations documented in the previous section are not inherent to EEG as a modality, but rather, they reflect the constraints of traditional feature engineering approaches. The fundamental challenge lies in learning representations that capture underlying neural dynamics while remaining invariant to nuisance factors, such as anatomical differences, electrode placement variations, and temporal drift, that have historically prevented generalization. Based on the traditional task-specific approaches, we have identified three essential requirements for generalizable EEG representations:

- Subject invariance with individual personalization. Models must learn features that capture general neural patterns while also dealing with variations based on the subject. The 19% accuracy drop seen in the cross-subject evaluation shows that current methods mix up anatomical differences with neural information. A strong representation should separate these factors: learning a shared feature space that encodes common neural dynamics while explicitly modeling individual differences through subject-conditional mechanisms. This allows the model to understand different subjects without losing information that might be important for a specific task.

- Temporal stability across sessions. Representations must remain stable despite session-to-session variability in electrode impedance, placement, and recording conditions. As demonstrated, EEG signals from the same subject can form distinct clusters across sessions, even under identical experimental protocols. Traditional features, being deterministic functions of the raw signal, inherit this instability directly. Learned representations, by contrast, can be trained to

extract invariant structure through exposure to diverse recording conditions, effectively learning to "see through" technical variations to the underlying neural processes.

- Task Transferability. Features should capture general properties of neural dynamics rather than task-specific patterns, enabling transfer across different BCI paradigms. Hierarchical representation learning addresses this by discovering multiple levels of abstraction: low-level features capture basic oscillatory patterns and spatial structures common across tasks, while higher-level features specialize for specific applications through fine-tuning. This mirrors successful approaches in computer vision, where ImageNet-pretrained models transfer effectively to diverse downstream tasks.

The limitations of task-specific approaches are similar to previous challenges in natural language processing and computer vision. However, representation learning has since transformed these two fields. Machine learning models have demonstrated their ability to learn general language representations applicable to various tasks. Similarly, computer vision models have demonstrated their ability to understand visual content in different contexts. Brain-computer interfaces are undergoing a similar evolution. The key insight is surprisingly simple. Instead of training specialized models for each task, we first learn general representations of neural signals that capture their fundamental structure. Then, we adapt these representations for specific applications. This approach inverts the traditional pipeline. Instead of engineering features for each task, we learn universal features that encode the rich information present in neural signals independently of any particular downstream application.

This paradigm is facilitated by the convergence of three important factors. First, large datasets such as the Temple University Hospital (TUH) Obeid & Picone (2016) corpus and the Mother of All BCI Benchmarks (MOABB) Aristimunha et al. (2025) provide the data heterogeneity needed for learning robust representations. Over thousands of subjects, over different recording paradigms, and over heterogeneous cognitive states, these datasets provide the heterogeneity needed for generalisa-

tion. Second, advances in computing power have allowed self-supervised learning approaches, which have proved to be the turning point for representation learning for other domains. Large models can now be learned on hundreds of hours of unlabelled electroencephalography recordings. Third, the remarkable success of foundation models in vision and language provides theoretical foundations and practical architectures that could be applied to fitting neural signals.

One model, pre-trained on diverse neural data, could be rapidly adapted for motor control, emotion recognition, medical diagnosis, or novel applications not yet envisioned—all while requiring minimal task-specific data for fine-tuning.

Parts III and IV of this dissertation develop two complementary approaches to learning generalizable EEG representations, addressing the requirements outlined above:

Part III: Continuous Representations through End-to-End Learning. Chapters 5 and 6 introduce deep neural architectures that learn continuous vector representations directly from raw EEG signals. Chapter 5 presents frequency-aware multi-band encoders for motor imagery classification. It demonstrates that learned spatiotemporal features outperform handcrafted approaches while maintaining interpretability through frequency decomposition. This architecture explicitly addresses subject and session variability by using learned subject embeddings.

Chapter 6 extends this framework to semantic decoding, introducing one of the first end-to-end systems for open-vocabulary EEG-to-text translation. The system achieves a BLEU-1 score of 42.75% across 30 subjects by aligning neural representations with pretrained language model spaces, demonstrating both subject generalization and cross-modal transfer. Using transformer encoders with subject-conditional layers provides a principled way to model universal and individual-specific neural patterns.

Part IV: Discrete Representations for Foundation Models. Chapter 7 introduces a different approach: vector quantization. This method converts continuous EEG streams into discrete ”brain tokens” drawn from a learned codebook. This process of discretization offers three significant bene-

fits. First, mapping signals to a finite vocabulary creates an explicit information bottleneck that discards irrelevant variations while retaining relevant task structure, achieving invariance across subjects and sessions. Second, discrete tokens allow for compositional reusability via token-level statistics and lightweight downstream classifiers. In this model, a pretrained tokenizer remains fixed while only small task-specific heads change. Third, discrete tokens serve as native inputs for modern large language models and multimodal transformers. This enables direct integration with architectures like BERT and GPT, opening pathways toward unified brain-language-vision models.

Part III

Learning Continuous Neural Representations

My mind seems to have become a kind of machine for grinding general laws out of large collections of facts.

Charles Darwin

5

Learning Spatiotemporal Representations from EEG Signals

Traditional machine learning and deep learning methods rely on handcrafted features, requiring domain expertise and often lacking robustness across subjects Casso et al. (2021); Angrisani et al. (2020); Zhang et al. (2019b); Lawhern et al. (2018); Ding et al. (2020); Schirrmeister et al. (2017). Many existing models process EEG signals holistically, which can overlook critical frequency-specific patterns necessary for MI classification. Additionally, deep learning models are highly sensitive to inter-subject variability, leading to reduced generalizability across different users and recording sessions. These challenges become even more pronounced when using dry electrodes, as their higher noise levels and unstable impedance further exacerbate variability and reduce the reliability of learned features. These limitations highlight the need for an improved approach that can better extract discriminative features while addressing frequency-specific information.

To address these limitations, in this work we propose WavEEGNet, a deep learning-based methodology for motor imagery classification using EEG signals recorded from a wearable BCI equipped with eight dry electrodes. WavEEGNet consists of multiple convolutional encoders (one per frequency band) and a residual Convolutional Neural Network (CNN) for feature fusion and classification.

5.1 Model Design and Experimental Setup

In this work, we propose a two-stage deep network for motor-imagery EEG. Stage 1 uses parallel encoders to capture temporal-spatial patterns in four frequency bands (delta, theta, alpha, and beta); Stage 2 integrates these via a residual CNN for final classification.

To assess its effectiveness, the proposed approach is evaluated on a very challenging EEG motor imagery dataset acquired using a wearable, non-invasive BCI device equipped with dry electrodes. Its performance is compared against both traditional machine learning methods and existing deep learning models.

5.1.1 EEG Dataset and Preprocessing

To comprehensively evaluate the performance of the proposed approach, we use a wearable MI-BCI-based dataset acquired with the *ab medica* Helmate device by Arpaia et al. (2022). Helmate is a non-invasive EEG Class IIA device certified according to the Medical Device Regulation (EU) 2017/745. It is equipped with 10 dry electrodes, plus reference and bias electrodes, positioned according to the international 10-20 EEG system Klem (1999). The electrode locations are FP1, FP2, Fz, Cz, C3, C4, O1, and O2, with reference and bias electrodes placed at AFz and FPz. The dataset consists of EEG recordings from 27 subjects performing two motor imagery tasks: imagining the movement of the left and right hand. Each participant completed five sessions, except for three participants (IDs 6, 17, and 27), who completed four. Each session included six runs, with each run consisting of 30 randomized trials, totaling 180 trials per session. EEG signals were recorded at a sampling rate of 512 Hz and transmitted via Bluetooth to a custom Simulink model. Evaluation follows an intra-subject protocol: for each participant, the data are partitioned into 70% training, 10% validation, and 20% testing sets, ensuring that every model is assessed exclusively on unseen trials from the same subject.

We preprocessed the MI-BCI dataset using the MNE-Python package Gramfort et al. (2013). First, EEG trials containing artifacts, as identified through visual inspection by the dataset authors, were removed. A band-pass filter (0.5 Hz – 30 Hz) was then applied to mitigate noise, signal artifacts, and power-line interference.

To remove physiological artifacts, such as eye blinks and muscle contractions, we applied Independent Component Analysis (ICA) Comon (1994). The signals were then resampled to 256 Hz. Next, the EEG signals were segmented into overlapping windows of 2.0 seconds with a 0.25-second overlap. For example, a 3-second segment (768 samples) was divided into five windows of 2.0 seconds (512 samples each). The final dataset contained 105,445 segments, with an average of 3,905 per subject (standard deviation: 330.75).

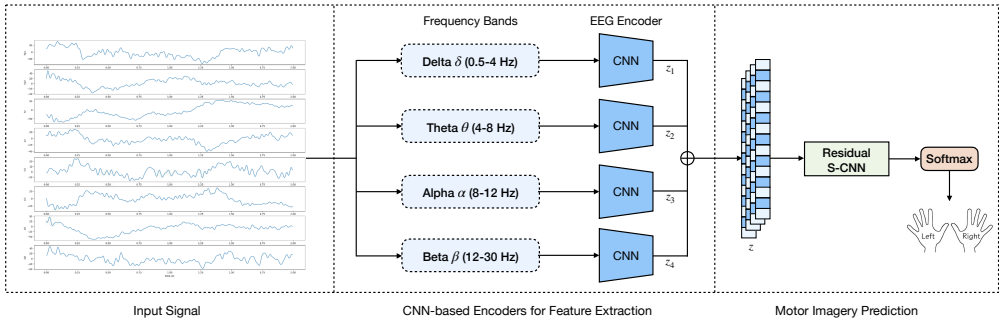


Figure 5.1: Overview of the proposed WavEEGNet architecture for motor imagery classification using multi-band EEG signals.

5.1.2 Network Architecture: WavEEGNet

We introduce WavEEGNet, a neural architecture for EEG-based BCIs that (1) facilitates motor imagery classification; (2) combines features from multiple EEG frequency bands; and (3) extracts neurophysiologically interpretable features.

WavEEGNet consists of multiple convolutional encoders (one per frequency band) and a residual CNN (RS-CNN) for feature fusion and classification. Figure 6.1 illustrates the architecture. The full description is reported in Table 5.1(a) and (b), respectively, for the CNN encoder, and the RS-CNN classification architecture.

Let $X \in \mathbb{R}^C$ be the raw EEG input, where C is the number of EEG channels. The input is first filtered into four frequency bands, delta (δ, 0.5-4 Hz), theta (θ, 4-8 Hz), alpha (α, 8-12 Hz), and beta (β, 12-30 Hz), each denoted as $X_\delta, X_\theta, X_\alpha$, and X_β, respectively. Each frequency band is then processed by a separate CNN encoder, denoted as $f_\delta(X_\delta), f_\theta(X_\theta), f_\alpha(X_\alpha)$, and $f_\beta(X_\beta)$. EEG signals are characterized by frequency-specific neural oscillations associated with different cognitive and motor functions. By independently processing the delta, theta, alpha, and beta frequency bands, WavEEGNet ensures that each oscillatory component contributes optimally to feature extraction without interference from other bands. Temporal convolution layers capture short- and long-range temporal

dependencies crucial for motor imagery classification, while spatial convolutions learn channel-wise relationships, improving robustness against noise. Furthermore, by integrating these features using a residual CNN, WavEEGNet enhances classification accuracy while preserving gradient flow during training.

In each CNN encoder, we perform two convolutional blocks in sequence. First, we use a spatial convolution and a temporal convolution. This two-step convolutional sequence allows to learn spatio-temporal features, enabling the efficient extraction of frequency-specific for each band. Both convolutions are linear, and do not use nonlinear activations. Along the feature map dimension, we apply (i) batch normalization, making the model more robust to changes in the distribution of inputs, (ii) average pooling, to reduce the sampling rate of the signal, and (iii) dropout, to help prevent over-fitting.

Then, we apply a separable convolution, which is a depth-wise convolution followed by a point-wise convolution. After each convolution there are a batch normalization layer and a ReLU (Rectified Linear Unit) activation function. The use of a separable convolution significantly reduces the number of fitting parameters, and separates the learning efforts. The depth-wise convolution focuses on how to summarize individual feature maps in time, while the point-wise convolution concentrates on how to optimally combine the feature maps. A flatten layer is then used. The output of each CNN encoder is a 1D feature vector, denoted as z_δ, z_θ, z_α, and z_β, respectively. The four 1D feature vectors are then stacked to form a 4D multi-channel representation of the EEG input, which is fed into the RS-CNN architecture, denoted as: $z = [z_\delta; z_\theta; z_\alpha; z_\beta]$.

The RS-CNN is based on the simplified CNN presented in Amrani et al. (2021) with the addition of residual learning connections He et al. (2016). We introduce residual connections (see Figure 6.1) to address the problem of vaninshing gradients, that can make difficult to learn and update the parameters of early results. Residual connections provide a shortcut path for the gradient flow during backpropagation, allowing for improved performance and faster convergence.

Table 5.1: Architectures of the proposed CNN encoder (left) and Residual S-CNN classification module (right).

Layer type	Layer characteristics	Output shape
Input		1 x 512 x 8
Spatial conv.	25 x 1, 40	40 x 488 x 8
Temporal conv.	1 x 8, 40	40 x 488 x 1
Batch normalization		40 x 488 x 1
Average pooling	75 x 1	40 x 28 x 1
Dropout		40 x 28 x 1
Depthwise separable conv.	28 x 1, 40	40 x 1 x 1
Batch normalization		40 x 1 x 1
ReLU activation		40 x 1 x 1
Pointwise separable conv.	1 x 1, 128	128 x 1 x 1
Batch normalization		128 x 1 x 1
ReLU activation		128 x 1 x 1
Flatten		128

(a) CNN encoder

Layer type	Layer characteristics	Output shape
Stacked Input		4 x 1 x 128
Convolution	1 x 5, 16	16 x 1 x 128
Batch normalization		16 x 1 x 128
#1 Convolution	1 x 5, 16	16 x 1 x 128
#1 Batch normalization		16 x 1 x 128
#1 Convolution	1 x 5, 16	16 x 1 x 128
#1 Batch normalization		16 x 1 x 128
#2 Convolution	1 x 5, 32, stride 2	32 x 1 x 64
#2 Batch normalization		32 x 1 x 64
#2 Convolution	1 x 5, 32	32 x 1 x 64
#2 Batch normalization		32 x 1 x 64
#3 Convolution	1 x 5, 64, stride 2	64 x 1 x 32
#3 Batch normalization		64 x 1 x 32
#3 Convolution	1 x 5, 64	64 x 1 x 32
#3 Batch normalization		64 x 1 x 32
Average pooling	1 x 10	64 x 1 x 23
Fully-connected	1472 x 512	512
ReLU activation		512
Fully-connected	512 x 2	2

(b) Residual S-CNN

RS-CNN consists of an initial convolutional block, followed by three residual blocks, an average pooling layer, and two fully connected layers. Each residual block is composed of a convolutional block and an addition operator, which sums the input of the residual block with the output of the residual block itself. A convolutional block is made up of two subsequent convolutional and batch normalization layers. It follows a fully connected layer with a ReLU activation function. Finally, the features are passed directly to a softmax classification with N units, N being the number of classes in the data, right and left, denoted as p.

The proposed model was trained by setting the cross-entropy loss function as follows:

$$L = -\frac{1}{M}\sum_{i=1}^{M}[t_i \log(p_i) + (1 - t_i)\log(1 - p_i)] \tag{5.1}$$

where M represents the data points, t_i the truth value taking 0 or 1, and p_i is the softmax probability for the i^{th} data point. The entire model, composed of four CNN encoders and the RS-CNN, was trained jointly through minimizing L in an end-to-end way, for a total of 200 epochs, saving the model weights that produced the lowest validation set loss. The learning rate was set to $1e-4$ with a drop factor of 10^{-1} after 100 epochs, and the batch size to 64. The model was implemented in Python, using the PyTorch framework.

5.1.3 Training Protocol and Evaluation Metrics

We compare the proposed approach with traditional machine and deep learning approaches, commonly used in EEG classification tasks. These are:

Hand-Crafted features. From the EEG signals, we have extracted a comprehensive set of hand-crafted features from three distinct domains: Time-Domain, Frequency-Domain, and Non-Linear features Stancin et al. (2021). For each channel x_i in the EEG input $X = x_{i_{i=1}}^{n=8}$, we computed the following features:

- Time-Domain features: the first three Hjorth parameters, including mobility, activity, and complexity, were computed based on the variance of the derivatives of the EEG signal;
- Frequency-Domain: the average bandpower of the EEG frequency bands in the delta (δ, 0.5-4 Hz), theta (θ, 4-8 Hz), alpha (α, 8-12 Hz), and beta (β, 12-30 Hz) ranges were calculated;
- Non-Linear features: the Petrosian fractal dimension (PFD), Higuchi fractal dimension (HFD),

Lempel-Ziv complexity (LZC), and Detrended Fluctuation Analysis (DFA) values were obtained.

The total number of features for the EEG input X is 88, which provided a comprehensive representation of the EEG signals in multiple dimensions and domains.

The performances of the Hand-Crafted features were evaluated using two Machine Learning algorithms: the Support Vector Machine (SVM) and the Multi-layer Perceptron (MLP). For the SVM classifier, a Radial Basis Function (RBF) kernel was utilized, with a regularization parameter (C) of 1.0 and a gamma coefficient calculated as $\frac{1}{m*\sigma^2(X)}$, where m denotes the number of features. The architecture of the MLP classifier consisted of three fully-connected layers with sizes 32, 64, and 2, respectively, and implemented Rectified Linear Unit (ReLu) activation functions, culminating in a final softmax layer.

Baseline models. EEGNet Lawhern et al. (2018) begins with a temporal convolution that learns frequency-selective filters, followed by a depthwise convolution that produces frequency-specific spatial filters. A separable convolution—depthwise for temporal summarization and pointwise for optimally combining feature maps—precedes the final classification layer.

Temporal Spatial Inception (TSception) Ding et al. (2020) consists of a temporal learner, a spatial learner, and a classifier. Multi-scale convolution kernels in the temporal and spatial learners capture representations at different time and space scales; the signal flows through the temporal learner, the spatial learner, and two fully connected layers to reach the target class.

Shallow ConvNet Schirrmeister et al. (2017) replicates the steps of FBCSP within a single network—temporal convolution, spatial convolution, squaring non-linearity, and mean pooling—so that all components can be trained jointly in an end-to-end fashion.

The proposed approach is evaluated and compared on the following metrics: Accuracy, Precision,

Recall, and F1-score. They are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.4}$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5.5}$$

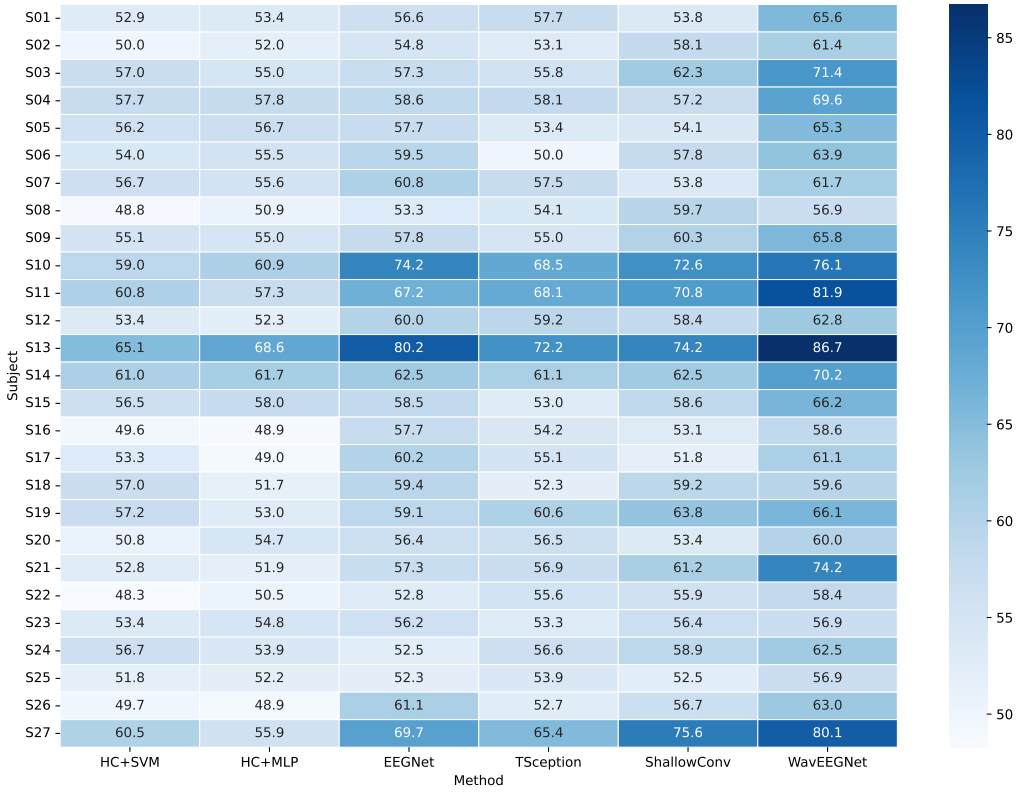
where True positive (TP) and false positive (FP) represent the number of correct and incorrect positive predictions made by the classification model, respectively. True negative (TN) and false negative (FN) represent the number of correct and incorrect negative predictions made by the model.

5.2 Experimental Results and Discussion

Table 5.2 presents the macro average accuracy for all methods and subjects in the EEG motor imagery dataset. This reflects the average classification accuracy per class across all sessions for each subject. Additionally, Table 5.3 provides the macro average values for accuracy, precision, recall, and F1-score, with each metric reported as the mean and standard deviation across session-specific test sets. The observed standard deviations are relatively high, indicating significant variability in classification performance across subjects and sessions.

The proposed method achieves an accuracy improvement of over 6% compared to traditional handcrafted feature-based approaches. This aligns with the performance of EEGNet, TSception, and Shallow ConvNet, collectively demonstrating that deep neural network-based feature extraction outperforms traditional feature engineering methods. The results suggest that learned neural representations better capture the underlying EEG patterns for motor imagery classification. Despite these improve-

Table 5.2: Heatmap of macro average accuracy (%) of all sessions by subject and method.



ments, the results also highlight the inherent challenges in classifying EEG motor imagery due to signal complexity and significant inter-subject and intra-subject variability. For example, classification accuracy varies substantially between subjects, with subject 8 achieving an average accuracy of 56.88%, whereas subject 11 exceeds 80%. These discrepancies suggest that some subjects produce more distinguishable EEG patterns for motor imagery tasks, whereas others exhibit higher noise levels or lower separability between classes.

A statistical evaluation using paired t-tests confirms the significant performance gains achieved by WavEEGNet in motor imagery classification. In all comparisons, the p-values are significantly lower than 0.05, indicating a statistically meaningful improvement. Additionally, WavEEGNet shows statistically significant differences compared to other methods, including HC + SVM, HC + MLP, EEG-

Table 5.3: Macro average metric of all subjects, accross the 27 subjects.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Hand Crafted features + SVM	55.01 (±4.52)	55.78 (±5.17)	54.72 (±4.69)	55.18 (±4.97)
Hand Crafted features + MLP	54.67 (±5.38)	55.69 (±5.9)	54.96 (±5.08)	55.3 (±5.44)
EEGNet	59.77 (±4.51)	60.62 (±4.61)	60.07 (±4.35)	60.34 (±4.44)
TSception	57.4 (±5.32)	57.6 (±5.29)	57.47 (±5.21)	57.53 (±5.25)
Shallow ConvNet	59.73 (±5.55)	60.79 (±6.31)	59.71 (±5.17)	60.19 (±5.6)
WavEEGNet	**66.03 (±5.74)**	**66.88 (±5.93)**	**65.95 (±5.66)**	**66.4 (±5.72)**

Net, TSception, and Shallow ConvNet, with p-values consistently below 1×10^{-6}. These results highlight the potential of WavEEGNet to enhance EEG-based motor imagery classification and contribute to the advancement of BCI technology.

5.3 Discussion and Limitations

While our results demonstrate clear improvements over established baselines using a realistic, consumer-grade EEG setup, some limitations remain. The evaluation relies on a single dataset, which may limit generalization. Future work will include validation on public benchmarks and heterogeneous acquisition settings.

Our study focuses on binary motor imagery classification, a widely adopted BCI task. Extending the approach to multi-class scenarios will be explored to assess scalability. Although rigorous statistical analysis was applied, we did not conduct cross-subject or cross-dataset validation. Future efforts will address this by incorporating subject-independent protocols and adaptation strategies.

Finally, while our frequency-specific architecture is motivated by neurophysiological principles that may enhance robustness, we plan to explicitly assess the model's resilience to noise, artifacts, and missing channels in real-world contexts.

In conclusion, this work presents WavEEGNet, a deep learning-based architecture designed for motor imagery classification using EEG signals recorded from a wearable brain-computer interface (BCI) with dry electrodes. The architecture independently processes multiple EEG frequency bands

by employing convolutional encoders to extract temporal and spatial features from four key frequency ranges (delta, theta, alpha, beta). These features are then combined using a residual convolutional neural network (RS-CNN) to enhance classification performance.

The effectiveness of WavEEGNet has been demonstrated through comparative evaluations with state-of-the-art approaches, including both traditional hand-crafted feature extraction methods and deep learning models. Experimental results confirm that WavEEGNet significantly improves motor imagery classification accuracy, even with the challenges posed by low-density, dry-electrode EEG signals. These findings are particularly relevant to the field of assistive technologies, where BCIs offer a promising solution for individuals with motor impairments. Despite the inherent challenges of signal complexity, noise, and inter-subject variability, the promising results achieved by WavEEGNet pave the way for future research focused on refining BCI applications to be more robust, effective, and widely accessible.

The limits of my language mean the limits of my world.

Ludwig Wittgenstein

6

Cross-Modal Alignment for EEG-to-Language Decoding

In this section, we present an end-to-end architecture for non-invasive brain recordings that uses pre-trained language models for open vocabulary EEG-to-text decoding. Firstly, our end-to-end deep learning architecture for open vocabulary EEG decoding incorporates a representation learning module for raw EEG encoding, a language modeling module based on BART Lewis et al. (2019), and a GPT-4 OpenAI (2023) refinement module, enhancing the comprehensibility of the generated sentences. The representation learning module includes a subject layer, which permits taking into account the subjectivity of EEG signals, and a multi-layer transformer encoder, which dynamically adapts to different lengths of word-level raw EEG signals. These allow the module to effectively extract latent brain representations, which are then aligned into language token embeddings. Secondly, we use the BERTScore Zhang et al. (2019a) in the evaluation, which incorporates semantic judgment at the sentence level, resulting in a more comprehensive evaluation that is closer to human perception. Thirdly, we conducted an ablation study to analyze and distinguish the contributions of each module within our proposal, including the specific impact of the GPT-4 refinement, providing valuable insights for future research work.

To demonstrate the efficacy of our approach, comprehensive evaluations are conducted on two publicly available datasets, ZuCo v1.0 and v2.0 Hollenstein et al. (2018, 2019), comprising EEG recordings from 30 subjects actively engaged in natural reading tasks. For a fair comparison, we keep the same teacher-forcing evaluation setting in EEG-To-Text Wang & Ji (2022) and DeWave Duan et al. (2023). The results achieved by our proposal, on previously unseen sentences, are a BLEU-1 score of 42.75%, a ROUGE-1-F Lin (2004) of 33.28%, and a BERTScore-F of 53.86%, surpassing the previous state-of-the-art results by 1.40%, 2.59%, and 3.20%, respectively.

Our code is available for public access at: `https://github.com/hamzaamrani/EEG-to-Text-Decoding`

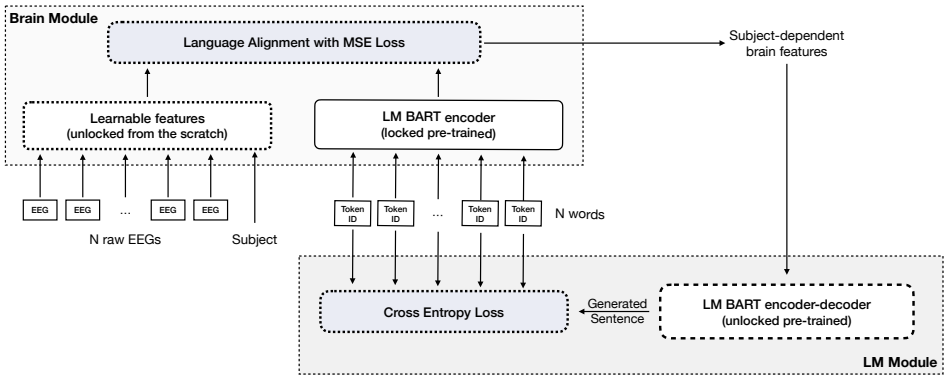


Figure 6.1: Overview of the proposed end-to-end architecture for open vocabulary EEG-to-Text decoding. Firstly, a sequence of word-level raw EEG signals is fed to the Brain module to extract deep-embedded representations for raw EEG encoding. Then, we use a Language Modeling (LM) module to generate EEG-to-Text sentences by leveraging the pre-trained language model BART. Dashed boxes correspond to the modules of the architecture that undergo training, while solid boxes represent the modules that remain untrained.

6.1 Architecture for EEG-Language Alignment

We aim to decode neural activity from a time series of high-dimensional brain signals recorded with non-invasive electroencephalography during the natural reading of English sentences. We first define the general task of open vocabulary EEG-to-Text decoding and then introduce the proposed end-to-end architecture.

6.1.1 Open Vocabulary EEG-to-Text Decoding

Let's define a sequence of word-level raw EEG signals as $X \in \mathbb{R}^{C \times T}$, with C the number of EEG channels and T the number of time steps. These EEG signals are a reflection of the recorded brain activity for a specific subject denoted as s, drawn from the set S consisting of distinct subjects. An EEG-decoding task is the task of predicting the corresponding text sentence Y in a Sequence-To-Sequence approach. Each text sentence Y is composed of English tokens $y_n \in \mathcal{V}$ from an open vocabulary $\mathcal{V}$. During the training phase, the EEG-subject-Text pairs can come from various subjects and various categories of reading materials.

Thus, a supervised EEG-to-Text decoding task consists in finding a decoding function $f : \{C \times T\} \times S \rightarrow \mathcal{V}$, such that f predicts Y given X and s. We denote by $\overline{Y} = f(X, s)$ the decoded/predicted text sentence from the brain signals.

Searching for f, the task is to maximize the probability of the decoded text sentence $\overline{Y}$:

$$p(\overline{Y}|X) = \prod_{n=1}^{N} p(\bar{y}_n \in \mathcal{V}|X, \bar{y}_{<n}) \tag{6.1}$$

where N is the length of the text sentence $\overline{Y}$, and $\bar{y}_n$ is the n-th token of $\overline{Y}$.

6.1.2 Proposed Architecture

The proposed architecture (refer to Section 6.4.1 for a detailed overview of the architecture) is composed of two main components: 1) a Brain module that implements a representation learning approach for EEG encoding; and 2) a Language Modeling module based on BART to produce EEG-to-Text sentences and on GPT-4 for sentence-level refinement. The training process is composed of two stages. An overview of the end-to-end architecture is presented in Figure 6.1, where dashed boxes correspond to the modules of the architecture that undergo training, while solid boxes represent the module that remains untrained. We start detailing the specifics of the training stages. Then we offer a more comprehensive breakdown of each module included in our architecture.

Training Stage 1

We initiate training with the Brain module: word-level EEG signals are aligned with word-tokens, as encoded by a locked, pre-trained BART Language Model, utilizing a Mean Square Error (MSE) Loss. This stage incorporates a learnable features module designed to account for EEG encoding and subjectivity. The outcome of this training stage yields EEG subject-dependent features. The alignment

procedure is done by mapping the learned EEG representation Z into the BART token embeddings $BART^{te}_{enc}$, using MSE regression loss $L_{MSE}(BART^{te}_{enc}, Z)$:

$$\min_{f_{brain}} \mathcal{L}_{MSE}(BART^{te}_{enc}, f_{brain}(X)) \tag{6.2}$$

Training Stage 2

Moving on, the subsequent step involves fine-tuning a pre-trained Language Model based on BART, aimed at generating word sequences through the utilization of a Cross-Entropy Loss. As in Wang & Ji (2022), we use the mapped embedded brain representation Z directly as initial word embeddings to feed into the pre-trained language model encoder-decoder BART Lewis et al. (2019). The high-level idea here is that we consider each embedded EEG representation as a word-level representation, and leverage a pre-trained language model to decode to real human language (English) like traditional machine translation tasks. Then, the last hidden states from the BART decoder are fed into a multi-layer perception (MLP) to generate English tokens $\bar{y}_n$ from the BART vocabulary $\mathcal{V}$.

During the training, the objective is to minimize the text reconstruction cross-entropy loss, defined as follows:

$$\mathcal{L}_{rec} = -\sum_{n=1}^{N} log\ p(\bar{y}_n \in \mathcal{V}) \tag{6.3}$$

Learnable Features Module

This module is included in the Brain module and it is used for extracting subject-dependent brain features from the raw EEG signals. Given a sequence of word-level raw EEG signals $X = \{x_0, x_1, ..., x_M\} \in \mathbb{R}^{C \times T}$ and the corresponding subject $s \in S$, we first use a deep neural network f_{brain} to get the latent subject-dependent brain representation $Z = \{z_0, z_1, ..., z_M\} = f_{brain}(X) \in \mathbb{R}^M$. This architecture (Figure 6.2) consists of (1) a learnable EEG feature block followed (2) by a subject layer to leverage

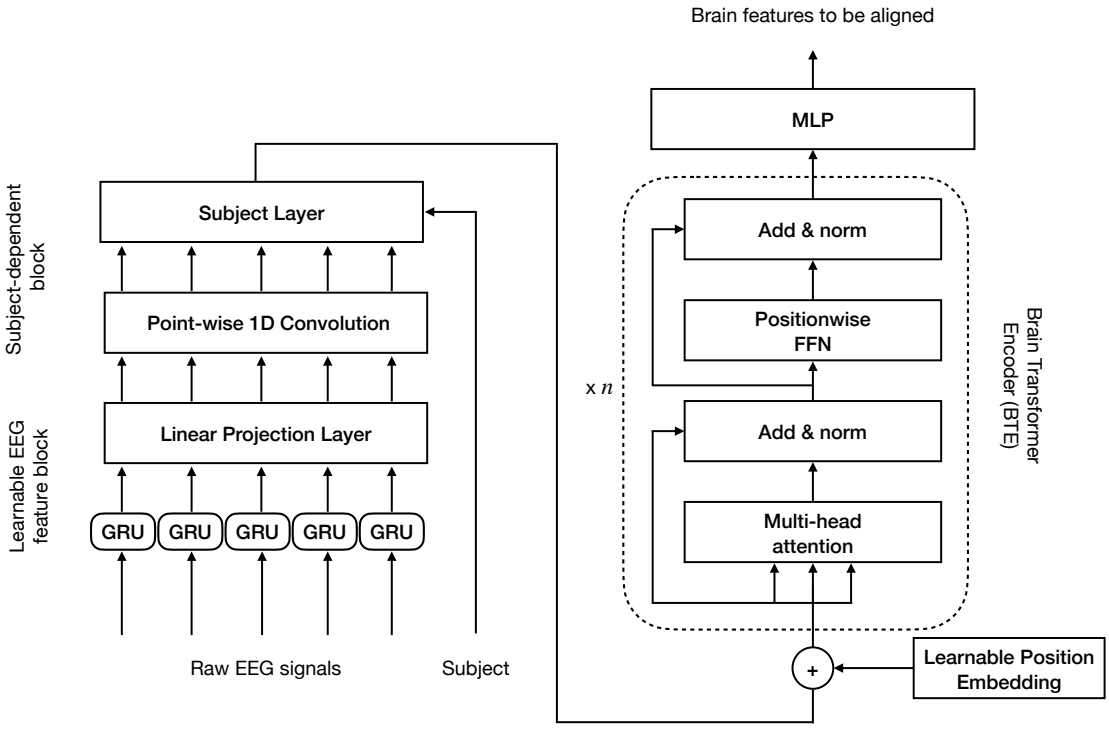


Figure 6.2: The Learnable features module consists of (1) a learnable EEG feature block, (2) a subject layer to leverage inter-subject variability, (3) a multi-layer transformer (Brain Transformer Encoder), and (4) an MLP.

inter-subject variability, which is input to (3) a multi-layer transformer encoder named *BTE* (Brain Transformer Encoder), and then to (4) a multi-layer perceptron.

The brain data is first fed to a bi-directional Gated Recurrent Unit (GRU) Cho et al. (2014) which reads the multi-time series input in both forward and backward directions to extract learnable EEG features. The use of GRU allows to dynamic address the different lengths of word-level raw EEG signals. We then apply a fully-connected layer to the concatenated forward and backward output. Similarly to Défossez et al. (2023b), we then add a 1x1 point-wise convolution (with a kernel size of 1) without activation and a number D of output channels. The use of a 1x1 pointwise convolution provides the dimensional congruence required for the subsequent application of the subject specific layer. This step ensures uniformity of feature map depth across different EEG signals, allowing our model

to adapt to and exploit variations between individual EEG signals for more personalised decoding. To leverage inter-subject variability, we learn a row vector $r_s \in \mathbb{R}^D$ for each subject $s \in S$ and apply it along the channel dimension. We then apply a multi-layer transformer encoder Vaswani et al. (2017) *BTE* with L layers, each with H attention heads and intermediate hidden dimension d_h. The inputs to the first layer BTE^0_{in} are produced using a weight matrix $W_{in} \in \mathbb{R}^{d_h \times l}$ and combined with a learnable 1D position embedding P Dosovitskiy et al. (2020), which is randomly initialized. Each layer applies self-attention with causal attention masking and a feed-forward layer to the input, with layer normalization Ba et al. (2016) and dropout Srivastava et al. (2014) being applied after. The outputs BTE^j_{out} of the j-th layer, become the inputs to the $(j+1)$-th layer. Then, the final outputs BTE^L_{out} are fed into a residual *MLP* network, composed of two fully connected layers, obtaining the latent brain representations z_m. As we will demonstrate subsequently in the ablation study, opting to process the raw EEG signals using a recurrent neural network, rather than directly handling pre-computed features as performed by Wang & Ji (2022), facilitates the extraction of subject-dependent nuances present in the brain recordings. These distinctive characteristics would otherwise remain entirely overlooked.

Sentence Refinement during Inference

During the inference phase, we propose the use of the pre-trained language model GPT-4 OpenAI (2023) via APIs on top of the generated text sentence $\overline{Y}$. It results in significant improvements in text comprehensibility, as well as a reduction in grammatical errors and repetitive words, enhancing the utility and effectiveness of the generated text sentence. The prompt used for the refinement is as follows:

As a text reconstructor, your task is to restore corrupted sentences to their original form while making minimum changes. You should adjust the spaces and punctuation marks as necessary. Do not introduce any additional information. If you are unable to reconstruct the text, respond with [False]. Reconstruct the following text: [text sentence $\overline{Y}$].

Table 6.1: ZuCo datasets statistics for each reading task. NR stands for Normal Reading, while TSR stands for Task-Specific-Reading.

Reading Task	#Sentences	#Train	#Val	#Test
NR v1.0	300	3,609	467	456
NR v2.0	349	2,645	343	350
TSR v1.0	407	4,456	522	601

6.2 Evaluation and Quantitative Results

We use Zurich Cognitive Language Processing Corpus (ZuCo) Hollenstein et al. (2018, 2019) datasets, which contain simultaneous electroencephalography and eye-tracking (ET) data recorded from natural reading tasks. The reading tasks include Normal Reading (NR) and Task-Specific Reading (TSR). The reading corpus of ZuCo are from movie reviews Socher et al. (2013) and Wikipedia articles. We used data from all the subjects in ZuCo v1.0 and v2.0 (12 and 18 respectively). For the EEG recordings, high-density data were recorded at a sampling rate of 500 Hz with a bandpass of 0.1 to 100 Hz, using a 128-channel EEG Geodesic Hydrocel system (Electrical Geodesics). The recording reference was set at electrode Cz. We follow Hollenstein et al. (2018, 2019) steps to perform data pre-processing on raw EEG signals, leading to 105 EEG channels from the scalp recordings.

In this section, we use concatenated sequences of word-level raw EEG signals, which were synchronized with ET fixations. We split each reading task's data (by unique sentences) into train, validation, and test (80%,10%,10%), as done by Wang & Ji (2022). The sentences in the test set are totally unseen. Table 6.1 shows the statistics of each reading task's data. Please refer to Section 6.4.2 for a detailed description of the electrodes used.

6.2.1 Training Details

Architecture Details

For the brain module, we set the GRU layer size to 512, and the fully connected layer to 1024. The 1d convolution maps to 64 channels and the 1d subject vector size is set to 64. The brain module comprises 171 million trainable parameters. The BTE has 12 layers and 8 attention heads, with an intermediate hidden dimension of 4096 and GELU activations Hendrycks & Gimpel (2016). The last hidden states of BTE are projected on a feature space of 1024. Then, we use the large version of BART, with 12 layers for the encoder and decoder, 8 attention heads, and an intermediate hidden dimension of 4096. BART includes a larger set of 561 million trainable parameters. For GPT-4, we use OpenAI's APIs and the model version *gpt-4*.

Optimization Settings

During training, we use the SGD optimizer with a cyclical learning rate set with $5e-7$ and $5e-5$ as initial and upper values to update model parameters. The batch size is set to 1 during the mapping between brain and word embeddings, then 8 during the training phase. During the alignment phase, a batch size of 1 provides greater control over the learning process, as each batch consists of a sentence of up to 56 words. For the training phase, a batch size of 8 has been empirically found to provide the best compromise between computational efficiency and model performance. The number of epochs is set to 25. During the training phase, we freeze the brain module weights. During inference, we use the model parameters on the best checkpoint based on the performance of the validation set.

For our architecture implementation, we use PyTorch[1] and Transformers (HuggingFace)[2] libraries. Both Stage1 and Stage2 were trained on a workstation equipped with Ubuntu 22.04, 32GB RAM

[1] https://github.com/pytorch/pytorch

[2] https://github.com/huggingface/transformers

and 2 Nvidia GeForce GTX 1070 with 8GB Memory. The average computation time per sentence is 0.0554s, with a standard deviation of 0.0110s.

Quantitative Metrics

In our experiments, we use BLEU and ROUGE metrics Papineni et al. (2002); Lin (2004) to measure the number of words shared by two sequences. However, the lexical congruence may not fully encapsulate semantic similarity due to lexical variations denoting similar meanings. To this end, we use BERTScore Zhang et al. (2019a), an approach that uses machine learning to capture the semantic similarity between two sequences by leveraging advanced language representations derived from the BERT model Devlin et al. (2018). BERTScore allows the integration of semantic similarity at the sentence level, leading to a more comprehensive evaluation that aligns with human perception.

For a fair comparison, these results keep the same teacher-forcing evaluation setting as EEG-To-Text Wang & Ji (2022) and DeWave Duan et al. (2023). This means that the decoding process eliminates accumulated errors and predicts the current token with the ground truth token from the last step.

Table 6.2: Open Vocabulary EEG-to-Text decoding model evaluation on ZuCo datasets. We compare our architecture with the current state-of-the-art by using three distinct metrics: BLEU-N ($N = 1, 2, 3, 4$), ROUGE-1 (Precision, Recall, and F1 scores), and BERTScore (Precision, Recall, and F1 scores). We also report ablations and the hypothetical upper limit for BART with fixation words when no errors are made to map EEG signals to token words. For a fair comparison, these results keep the same teacher-forcing evaluation setting as EEG-To-Text Wang & Ji (2022) and DeWave Duan et al. (2023). **Bold** numbers indicate the first best result, Underline numbers indicate the second best result.

Method	**BLEU-N (%) ↑**				**ROUGE-1 (%) ↑**			**BERTScore (%) ↑**		
	N=1	N=2	N=3	N=4	R	P	F	P	R	F
EEG-To-Text Wang & Ji (2022)	40.1	23.1	12.5	6.8	28.8	31.7	30.1	48.84	52.71	50.66
DeWave Duan et al. (2023)	41.35	24.15	13.92	8.22	28.82	33.71	30.69	-	-	-
Our Architecture	**42.75**	**25.90**	**15.66**	**9.56**	**30.60**	**36.71**	**33.28**	52.62	55.26	53.86
w GPT-4	40.87	24.43	14.53	8.82	30.40	35.50	32.61	**54.58**	**56.30**	**55.34**
w/o subject layer	41.51	24.41	14.31	8.38	29.22	35.40	31.92	51.09	53.93	52.43
w/o language alignment	41.30	24.50	14.14	8.40	29.16	35.76	32.00	50.82	53.62	52.16
w/o BTE	35.51	20.51	12.61	8.98	25.62	26.38	25.83	46.44	50.52	48.34
w/o BART finetuning	28.50	14.35	7.01	3.38	21.32	23.07	22.03	39.67	47.90	43.13
BART with fixation words	72.45	62.16	53.80	46.84	67.16	75.25	70.65	66.72	74.47	69.89

Table 6.3: Open Vocabulary EEG-to-Text decoding examples on ZuCo unseen test sentences. We report both predictions from our model, with and without GPT-4 sentence refinement. (1-3) are in NR v1.0, v2.0. (4) is in SR v1.0. **Bold** means exact match, *Italic* indicates semantic similarity. Underline denotes error match.

(1)	Ground truth	He is a prominent member of the Bush family, the younger brother of President George W. Bush...
	EEG-To-Text Wang & Ji (2022)	was a former **member of** the *American family*, and son brother of **President George W. Bush**...
	Prediction	was the member **member** of the *American family*. and **younger brother of President George W. Bush**
	Prediction + GPT-4	He was a **member of the American Bush family**, **brother of President George W. Bush**...
(2)	Ground truth	Raymond Arrieta (born March 26, 1965 in San Juan, Puerto Rico) is considered by many to be one of Puerto Rico's greatest comedians.
	EEG-To-Text Wang & Ji (2022)	mond wasaga,19 in 17, 18) New Francisco, **Puerto Rico)** is a one many to be the of the *Rico's* **greatest** poets.
	Prediction	mond wasaga **(born** April 17, 1946) New Francisco, **Puerto Rico)** is a one many to be the of the *Rico's* **most** *artists*.
	Prediction + GPT-4	Ramon Wasaga **(born** April 17, 1946, in New Francisco, **Puerto Rico)** is one of the many to be considered as one of **the *most prominent artists* of Puerto Rico**.
(3)	Ground truth	Following the 1980 presidential election, Bush and his family moved to Miami-Dade County, Florida.
	EEG-To-Text Wang & Ji (2022)	the deaths **election**, the was *his wife* moved to California,**Dade County, Florida**
	Prediction	the wars **election**, **Bush** was *his wife* moved to Florida,**Dade County, Florida**.
	Prediction + GPT-4	*After* the war's **election**, **Bush** and *his wife* moved to **Dade County, Florida**.
(4)	Ground truth	It's not a particularly good film, but neither is it a monsterous one.
	EEG-To-Text Wang & Ji (2022)	was a a *bad* **good** story, but it is it *bad bad.* one.
	Prediction	's a a *bad* **good movie**, but it is it *bad bad.* one.
	Prediction + GPT-4	It's a *bad* **good movie**, but is it a *bad* one.

6.2.2 Experimental Results

We compared our architecture with the current state-of-the-art models by Wang & Ji (2022); Duan et al. (2023). As shown in Table 6.2, our proposal achieves a BLEU-1 score of 42.75%, a ROUGE-1-F of 33.28%, and a BERTScore-F of 53.86%, showing an improvement over the state-of-the-art by 1.40%, 2.59%, and 3.2%, respectively. For larger *n*-grams evaluation, we obtain BLEU-{2,3,4} scores of 25.90%, 15.66%, and 9.56% respectively, leading to an increase of 1.75%, 1.74%, and 1.34%. Our decoding embeddings resulted in higher performance for each metric, demonstrating the positive impact of learning embedded EEG representations and exploiting intersubject variability. In Section 6.4.3 we report the obtained results of our architecture for each subject. The results show a significant difference between v1.0 and v2.0 participants. On average, v2.0 participants outperform v1.0 participants by 19.64%, 42.61%, and 11.83% for BLEU-1, ROUGE1-F, and BERTScore-F respectively.

In addition to numerical results, we report decoding examples of generated EEG-to-Text sentences compared to the ground truth and the state of the art, with and without GPT-4 sentence refinement

(Table 6.3). We observe that our model is sometimes able to precisely capture named entities that do not exist in the training set. "George W. Bush" in (1) and "Puerto Rico" in (2) are correctly decoded, while "presidential election" in (3) is incorrectly decoded. Compared to Wang & Ji (2022), our model results in significant improvements in text comprehensibility, as well as a reduction in grammatical errors and repetitive words, as shown in example (4). Please refer to Section 6.4.4 to see additional decoding examples of generated EEG-to-Text sentences.

The complexity of open vocabulary EEG decoding tasks arises from the high dimensionality, intersubjectivity, and variability of EEG data, coupled with the intrinsic difficulties associated with the language decoding capabilities of AI-based language models. Our improvements represent significant progress in overcoming these multiple challenges and suggest a promising direction for future research in non-invasive brain decoding.

Embedding Visualization

We provide a visual comparison via t-distributed stochastic neighbor embedding (t-SNE) between the pre-calculated EEG features (Figure 6.3 (left)) as used by Wang & Ji (2022), and EEG embedded representations obtained by the proposed Brain module (Figure 6.3 (right)). In the visualization, different colors indicate different subjects to delineate the variation between individuals. Each dot in the scatter plot corresponds to an EEG representation of a sentence. In particular, a red triangle is used to specifically highlight the EEG-embedded representation of the sentence *"With his interest in race cars, he formed a second company, the Henry Ford Company."*, allowing for a direct visual comparison of how this particular sentence is represented differently in the two methods. Our learned EEG representations of sentences show more pronounced clustering by subject compared to the pre-calculated EEG features, highlighting the ability of our brain module to capture subject-specific variations in EEG data. Adaptation to individual variability allows for a personalized approach to processing the output of the brain module for each subject. This methodology exploits the intrinsic EEG differences

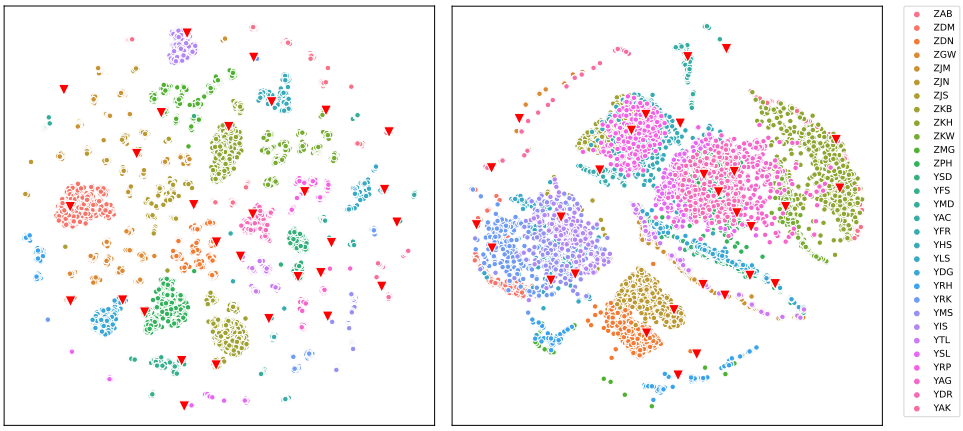


Figure 6.3: t-SNE visualization of EEG embedded representations of sentences in the training set, which are (left) original EEG representations and (right) generated by the Brain module of our architecture. Distinct colors mean different subjects. Each dot represents a sentence. The red triangle represents the EEG embedded representations corresponding to the same sentence *"With his interest in race cars, he formed a second company, the Henry Ford Company"*.

between individuals, thus improving the robustness and accuracy of the decoding processes across subjects.

6.2.3 Ablations

Our ablations highlight the importance of (1) GPT-4 sentence refinement, (2) the subject layer, (3) the language alignment, (4) the use of the Brain Transformer Encoder, and (5) the BART finetuning (Table 6.2). First, we evaluated the effect of GPT-4 sentence refinement on the performance of our model. Unlike the base architecture, which focuses primarily on syntactic accuracy, the integration of GPT-4 improves the semantic coherence of the generated sentences. This difference is evident in the comparative analysis, where the model incorporating GPT-4 shows improved semantic accuracy. For example, while syntax-related metrics such as BLEU-1 and BLEU-2 show relatively small changes, semantic-sensitive measures such as BERTScore show more substantial improvements. This underscores GPT-4's ability to refine sentences in ways that better capture intended meanings and nuances, rather than simply improving surface-level grammatical structure. Second, a model trained to gener-

ate EEG-to-Text sentences without the use of the subject layer achieves lower decoding accuracy on average across datasets, that is, about $1-1.5\%$ lower than our model. While modest, these scores show the positive effect of leveraging inter-subject variability. Third, we show the effect of using language alignment with MSE. The results show small differences, especially in the BLEU and ROUGE scores. For BERTScore we see small improvements. Sentence generation without the Brain Transformer Encoder shows a significant drop in performance compared to our model. For BLEU-1 the decrease is 7.24%, while for BLEU-2 is 5.39%. While, ROUGE-1-F and BERTScore-F lose 7.45% and 5.52%, respectively. We verified that the Brain Transformer Encoder provides higher decoding performances. Finally, to test whether our model effectively leverages the pre trained BART model, we trained it without fine-tuning the BART model weights. As reported, decoding performance decreases notably up to 14.25%. This loss significantly confirms the use of fine-tuning on the BART model.

Then, we also show the hypothetical upper limit for EEG-to-Text decoding when no errors are made to map EEG signals to token words. Separately from our model, we fine-tuned BART on only Eye-Tracking fixations words without considering the raw EEG signals to reconstruct the original text sentence. It outperforms our proposed architecture by about 30% in terms of BLEU-1, 37% in terms of ROUGE-1-F, and 15% in terms of BERTScore-F. The obtained results reveal the existence of two challenges within the EEG-to-Text decoding task. The initial challenge pertains to the model's capacity to establish a dependable EEG-feature representation for the word tokens. The subsequent challenge involves the faithful reconstruction of the sentence. This experiment highlights that, between these two challenges, the foremost one is undoubtedly the ability to discern an efficacious representation of the EEG signals. This observation thereby points towards the direction of future research efforts.

6.3 Discussion and Future Works

In this section, we present an end-to-end architecture for open vocabulary EEG-to-Text decoding task. By leveraging a subject-dependent representation learning module, a pre-trained BART language model, and a GPT-4 sentence refinement module, this study offers a comprehensive solution that not only enhances decoding performance but also delves into the human comprehensibility of the decoded output. The incorporation of the BERTScore as an evaluation metric has enabled a more holistic assessment, capturing not only syntactic accuracy but also taking into account human understanding at the sentence level. Moreover, the conducted ablation study permitted us to understand the contribution to the proposed architecture of each component. This in-depth analysis not only validates the efficacy of each module but also provides a roadmap for further research, guiding the development of refined and optimized approaches in the future.

The empirical validation on two publicly available datasets demonstrates the effectiveness of the proposed architecture, achieving a BLEU-1 score of 42.75%, a ROUGE-1-F of 33.28%, and a BERTScore-F of 53.86%. Our results show that the use of raw EEG signals leads to improved results, demonstrating the effectiveness of modern representational learning approaches in neuroscience.

In summary, this research not only fills critical voids in the EEG decoding landscape but also shows the way for future investigations. By combining advanced neural network architectures with sophisticated evaluation methodologies, the study pushes the boundaries of EEG-to-text decoding and encourages continued innovation in the pursuit of more accurate and human-aligned results. One future direction is to improve the quality of the generated embedded representations by taking into account inter-subject variability, so to increase the ability of the model to generalize across individuals. Furthermore, ethical considerations need to be at the forefront as we move forward. Ensuring privacy, establishing clear guidelines for consent, and considering the potential long-term effects of this technology on users is critical.

6.4 Supplementary Details

6.4.1 Architecture

A detailed overview of the architecture is given in Figure 6.4. It is composed of two main components: 1) a Brain module that implements a representation learning approach for EEG encoding; and 2) a Language Modeling module based on BART to produce EEG-to-Text sentences and on GPT-4 for sentence-level refinement.

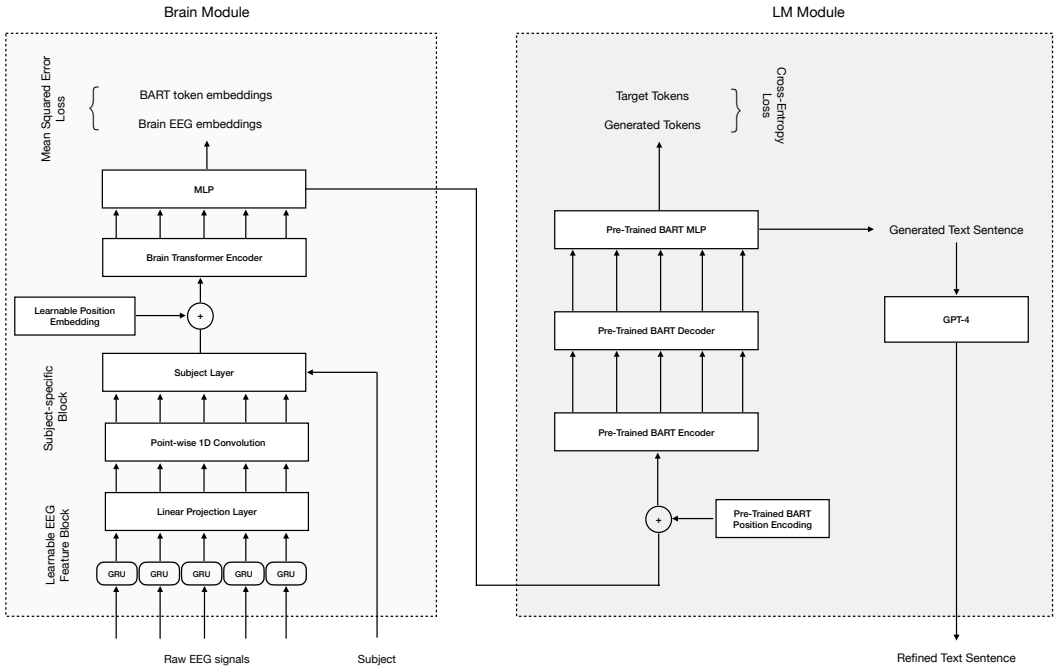


Figure 6.4: End-to-end architecture for open vocabulary EEG-to-Text decoding.

6.4.2 Dataset EEG electrodes

In ZuCo dataset Hollenstein et al. (2018, 2019), we follow Hollenstein et al. (2018, 2019) steps to perform data pre-processing on raw EEG signals, leading to 105 EEG channels from the scalp recordings.

It follows the full list of EEG channels: *E2, E3, E4, E5, E6, E7, E9, E10, E11, E12, E13, E15, E16, E18, E19, E20, E22, E23, E24, E26, E27, E28, E29, E30, E31, E33, E34, E35, E36, E37, E38, E39, E40, E41, E42, E43, E44, E45, E46, E47, E50, E51, E52, E53, E54, E55, E57, E58, E59, E60, E61, E62, E64, E65, E66, E67, E69, E70, E71, E72, E74, E75, E76, E77, E78, E79, E80, E82, E83, E84, E85, E86, E87, E89, E90, E91, E92, E93, E95, E96, E97, E98, E100, E101, E102, E103, E104, E105, E106, E108, E109, E110, E111, E112, E114, E115, E116, E117, E118, E120, E121, E122, E123, E124, Cz.*

In this work, the Cz EEG channel has been removed as it consists of all zeros.

6.4.3 Decoding accuracy results by subject

We report open vocabulary EEG-to-Text decoding results for each subject (see Table 6.4). The results show a significant difference between subjects from the v1.0 and v2.0 of the dataset. The v2.0 results achieve a BLEU-1 score of 47.13%, a ROUGE-1-F of 40.16%, and a BERTScore-F of 57.35%, while the v1.0 results obtain a BLEU-1 score of 39.39%, a ROUGE-1-F of 28.16%, and a BERTScore-F of 51.28%, so leading to an increment of 19.64%, 42.61% and 11.83% respectively.

Table 6.4: Open Vocabulary EEG-to-Text decoding model evaluation on ZuCo datasets by each subject.

Subject	ZuCo	BLEU-N (%) ↑				ROUGE-1 (%) ↑			BERTScore (%) ↑		
		N=1	N=2	N=3	N=4	R	P	F	P	R	F
ZAB	v1.0	39.38	22.11	11.92	6.61	25.94	30.92	28.11	49.88	52.62	51.16
ZDM	v1.0	39.45	22.24	12.02	6.67	25.93	30.94	28.11	50.00	52.73	51.28
ZDN	v1.0	39.06	21.93	11.81	6.63	26.12	31.25	28.35	49.80	52.45	51.04
ZGW	v1.0	39.79	22.57	12.27	6.92	26.08	30.98	28.22	50.34	53.07	51.62
ZJM	v1.0	39.27	21.99	11.97	6.67	25.94	30.96	28.12	49.73	52.46	51.00
ZJN	v1.0	39.76	22.52	12.49	7.05	26.51	31.48	28.68	50.37	53.07	51.64
ZJS	v1.0	39.22	22.49	12.23	6.82	25.66	30.29	27.69	50.47	53.23	51.76
ZKB	v1.0	39.38	22.11	11.92	6.61	25.94	30.92	28.11	49.88	52.62	51.16
ZKH	v1.0	39.32	22.01	11.86	6.60	26.00	31.00	28.18	49.86	52.60	51.14
ZKW	v1.0	39.38	22.11	11.92	6.61	25.94	30.92	28.11	49.88	52.62	51.16
ZMG	v1.0	39.29	22.22	12.02	6.70	25.95	30.93	28.12	49.93	52.68	51.22
ZPH	v1.0	39.38	22.11	11.92	6.61	25.94	30.92	28.11	49.88	52.62	51.16
YSD	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YFS	v2.0	47.09	30.87	20.29	13.15	36.76	44.65	40.24	56.21	58.68	57.37
YMD	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YAC	v2.0	46.88	30.25	19.92	12.90	36.59	44.62	40.14	56.22	58.52	57.30
YFR	v2.0	45.82	29.23	19.09	12.21	35.91	42.64	38.91	56.51	59.13	57.74
YHS	v2.0	47.22	30.55	20.00	12.92	36.80	44.41	40.16	56.06	58.60	57.26
YLS	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YDG	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YRH	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YRK	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YMS	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YIS	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YTL	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YSL	v2.0	47.52	31.00	20.34	13.20	37.23	44.98	40.65	56.54	59.02	57.71
YRP	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YAG	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
YDR	v2.0	47.16	30.63	20.23	13.17	37.00	44.70	40.40	56.31	58.74	57.45
YAK	v2.0	47.22	30.64	20.04	12.94	36.80	44.47	40.19	56.10	58.63	57.29
Average	v1.0	$39.39_{\pm 0.20}$	$22.20_{\pm 0.21}$	$12.03_{\pm 0.19}$	$6.71_{\pm 0.14}$	$26.00_{\pm 0.19}$	$30.96_{\pm 0.26}$	$28.16_{\pm 0.21}$	$50.00_{\pm 0.24}$	$52.73_{\pm 0.24}$	$51.28_{\pm 0.24}$
	v2.0	$47.13_{\pm 0.34}$	$30.57_{\pm 0.35}$	$20.02_{\pm 0.25}$	$12.93_{\pm 0.20}$	$36.77_{\pm 0.24}$	$44.42_{\pm 0.45}$	$40.16_{\pm 0.32}$	$56.17_{\pm 0.14}$	$58.68_{\pm 0.15}$	$57.35_{\pm 0.14}$
	v1.0 + v2.0	$44.03_{\pm 3.80}$	$27.22_{\pm 4.11}$	$16.82_{\pm 3.92}$	$10.44_{\pm 3.06}$	$32.46_{\pm 5.28}$	$39.04_{\pm 6.61}$	$35.36_{\pm 5.88}$	$53.70_{\pm 3.03}$	$56.30_{\pm 2.92}$	$54.92_{\pm 2.98}$

6.4.4 Decoding Examples

We report additional decoding examples of generated EEG-to-Text sentences (see Table 6.5), with and without GPT-4 sentence refinement. The prompt used for the GPT-4 sentence refinement is as follows:

As a text reconstructor, your task is to restore corrupted sentences to their original form while making minimum changes. You should adjust the spaces and punctuation marks as necessary. Do not introduce any additional information. If you are unable to reconstruct the text, respond with [False]. Reconstruct the following text: [text sentence $\overline{Y}$].

Table 6.5: Open Vocabulary EEG-to-Text decoding examples on ZuCo unseen test sentences, with and without GPT-4 sentence refinement.

(1)	Ground truth	An amateurish, quasi-improvised acting exercise shot on ugly digital video.
	Prediction	interesting actor, un-religiousprovised film performance, through a, video.
	Prediction + GPT-4	Interesting actor, un-religious, improvised film performance, through a video.
(2)	Ground truth	Viewed as a comedy, a romance, a fairy tale, or a drama, there's nothing remotely triumphant about this motion picture.
	Prediction	the from a kind of it satire, and love tale, and a love, it's a quite funny about it film picture.
	Prediction + GPT-4	From a kind of satire, it's a love tale and quite a funny film picture about love.
(3)	Ground truth	It's solid and affecting and exactly as thought-provoking as it should be.
	Prediction	's a, well. it what it-provoking as the sounds be.
	Prediction + GPT-4	Well, it's as provoking as it sounds, what a be.
(4)	Ground truth	It's a head-turner – thoughtfully written, beautifully read and, finally, deeply humanizing.
	Prediction	s a greatyscing, a to crafted, well acted, well most, a moving..
	Prediction + GPT-4	It's a great, most moving, well-crafted and well-acted scene.
(5)	Ground truth	"The Kid Stays in the Picture" is a great story, terrifically told by the man who wrote it but this Cliff Notes edition is a cheat.
	Prediction	The movie"ays in the House" is a film movie about andally funny by a young who wrote it. also ish version is a little.
	Prediction + GPT-4	"The movie, 'Days in the House', is a film about a young man who wrote it. It's also randomly funny. The British version is a little different.
(6)	Ground truth	Fans of the TV series will be disappointed, and everyone else will be slightly bored.
	Prediction	of the film series will recognize familiar to but the will will be happy disappointed.
	Prediction + GPT-4	of the film series will recognize familiar to but the will be happy disappointed.
(7)	Ground truth	Wedding feels a bit anachronistic
	Prediction	alting bells like little likeachronistic,
	Prediction + GPT-4	alting bells like little likeachronistic.
(8)	Ground truth	But what's nice is that there's a casual intelligence that permeates the script.
	Prediction	he's most about that it's a sense, to'sates the film.
	Prediction + GPT-4	He's most about that. It's a sense to states the film.
(9)	Ground truth	An important movie, a reminder of the power of film to move us and to make us examine our values.
	Prediction	interesting part about but must of the importance of the to shape people. of make us think our lives.
	Prediction + GPT-4	interesting part about but must of the importance of the to shape people. of make us think our lives.
(10)	Ground truth	Jeb Bush was born in Midland, Texas, where his father was running an oil drilling company.
	Prediction	uan Bush was born in Newland, Texas, and his father was a a insurance company company.
	Prediction + GPT-4	Juan Bush was born in Newland, Texas, and his father was an insurance company owner.

Part IV

Discrete Representations and Foundation Models

Nature uses only the longest threads to weave her patterns, so each small piece of her fabric reveals the organization of the entire tapestry.

Richard P. Feynman

7

Vector Quantized EEG Representations

7.1 Motivation: Why Discretize Neural Signals?

Non-invasive neural signals, such as EEG, are high-dimensional, non-stationary, and subject to significant variability between and within subjects. Although continuous encoders can learn powerful spatiotemporal features, their latent spaces often vary across sessions and are difficult to align between datasets. Transitioning from continuous latents to a finite set of discrete codes provides several advantages, particularly for EEG.

First, discretization creates an explicit information bottleneck: short segments of multichannel time series are mapped to a small vocabulary of latent symbols (code indices). This reduces unwanted variation (e.g., minor shifts in electrode placement, impedance, or state) by assigning nearby points on the signal manifold to the same cell (Voronoi region). This increases within-cluster stability while preserving task-relevant distinctions. Here the codebook size controls the trade-off between fidelity and invariance: larger codebooks capture finer structures, while smaller codebooks encourage more general abstraction.

Second, discrete latents promote compositionality and reuse. If recurring spatiotemporal patterns are consistently mapped to the same code, then longer sequences can be analyzed as arrangements of a small set of codes. This is valuable for interpretability, we can visualize the average waveform or spectrum per code, and for transferability, the same codebook can be reused across subjects and tasks.

Third, discrete tokens enable token-level objectives (masking, autoregression, and contrastive prediction) and lightweight downstream models. Once an EEG stream is tokenized, simple histograms, n-grams, or temporal pooling over code embeddings provide strong baselines for classification, often with better sample complexity than training continuous models from scratch.

Finally, discretization is computationally attractive because encoding compresses continuous signals by orders of magnitude. Models that operate on token sequences have predictable memory and latency, facilitating real-time BCI.

7.2 Vector Quantized Variational Autoencoder for EEG

7.2.1 Problem formulation

We consider a multi–channel EEG recording as $X \in \mathbb{R}^{C\times T}$, where T is the total timestamps. Let $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$ be the universal set of channels in the international 10-20 system. Each recording uses a subset $\mathcal{C}_X \subseteq \mathcal{C}$. Throughout, we denote $C = |\mathcal{C}_X|$ and order the rows of X according to $\mathcal{C}_X$.

For learning, X is divided into segments of fixed duration t with hop s. The number of samples is $N = \left\lfloor \frac{T-t}{s} \right\rfloor + 1$, and the i-th sample is $x^{(i)} \in \mathbb{R}^{C\times t}, \qquad i = 1, \ldots, N.$

Within a sample $x \in \mathbb{R}^{C\times t}$, each channel is then partitioned into contiguous patches of length w, obtaining

$$x = \left\{ x_{c_j,k} \in \mathbb{R}^w \,\middle|\, j = 1, \ldots, C,\ k = 1, \ldots, \left\lfloor \tfrac{t}{w} \right\rfloor \right\}, |x| = C\left\lfloor \tfrac{t}{w} \right\rfloor.$$

Our goal is to map each patch to a discrete latent symbol from a finite codebook and to reconstruct the spectral representation of the original signal from these symbols. Formally, an encoder $f_\theta : \mathbb{R}^w \to \mathbb{R}^D$ produces embeddings

$$\mathbf{h}_{c,k} = f_\theta(x_{c,k}) \in \mathbb{R}^D.$$

A vector-quantizer with codebook $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\} \subset \mathbb{R}^D$ assigns each embedding to its nearest code. The discrete sequence $z(x) = \{z_{c,k}\}$ constitutes the tokenized representation of x. A decoder g_φ maps the quantized vectors back to the spectral domain to obtain per–patch reconstructions. The training objective and the codebook update rules are described in the following.

7.2.2 Neural Tokenizer and Fourier Spectrum Targets

Neural tokenizer. Let $p = \{p_i\}_{i=1}^N$ denote the set of patch embeddings produced by the encoder for a segment x, with $N = C\lfloor t/w \rfloor$ (one embedding per channel–patch). We define a codebook

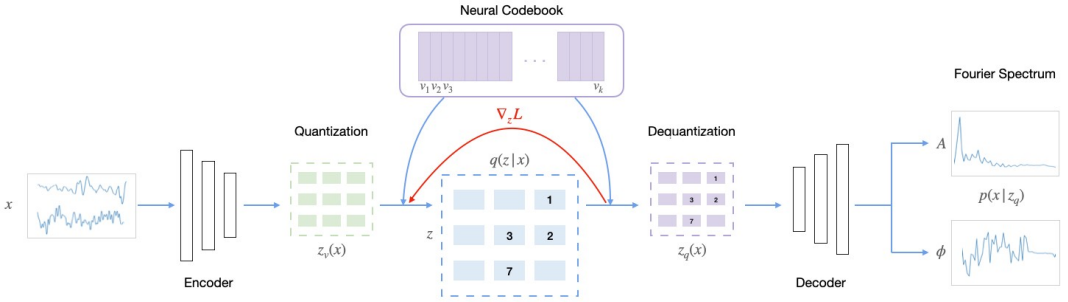


Figure 7.1: Neural tokenization architecture employing vector quantization. The encoder maps input signal x to latent space $z_e(x)$, which is discretized through a neural codebook. The decoder reconstructs the Fourier spectrum from quantized tokens $z_q(x)$, with $\mathcal{L}_{VQ}$ loss ensuring alignment between continuous and discrete representations.

$\mathcal{E} = \{v_j\}_{j=1}^{K}$, $v_j \in \mathbb{R}^D$. Quantization is performed by nearest neighbor search in cosine space via ℓ_2 normalization:

$$z_i \;=\; \arg\min_{j\in\{1,\dots,K\}} \big\| \ell_2(p_i) - \ell_2(v_j) \big\|_2, \qquad \hat{v}_i \;=\; v_{z_i},$$

where $\ell_2(\cdot)$ denotes normalization to unit Euclidean norm. This is equivalent to selecting the code of maximal cosine similarity and has been observed to improve code utilization Peng et al. (2022).

FOURIER SPECTRUM PREDICTION. EEG patches have low SNR and nonlinear artifacts, making direct regression of raw samples brittle. As Jiang et al. (2024), we predict the Discrete Fourier Transform (DFT) of each patch. For a channel–patch $x_{c,k} \in \mathbb{R}^w$, the DFT at frequency index m is

$$\tilde{x}^m_{c,k} \;=\; \sum_{n=1}^{N} x_{c,k}[n] \, \exp\Big(-\frac{2\pi j}{N} mn \Big).$$

Let $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote real and imaginary parts. Amplitude and phase are computed as

$$A^m_{c,k} \;=\; \sqrt{\mathrm{Re}(\tilde{x}^m_{c,k})^2 + \mathrm{Im}(\tilde{x}^m_{c,k})^2}, \qquad \phi^m_{c,k} \;=\; \arctan\Big(\frac{\mathrm{Im}(\tilde{x}^m_{c,k})}{\mathrm{Re}(\tilde{x}^m_{c,k})}\Big).$$

Within each segment we apply z-score normalization to $\{A^m_{c,k}\}$ and $\{\phi^m_{c,k}\}$ for stable optimization.

The overall architecture is shown in Figure 7.1.

7.2.3 Learning Objective and Codebook Updates

The normalized discrete embeddings $\{\ell_2(\hat{v}_i)\}_{i=1}^{N}$ are passed to a lightweight decoder composed of several Transformer blocks. Outputs are temporally aggregated by average pooling, followed by two prediction heads that regress amplitude and phase, obtaining $\{\hat{A}^m_{c,k}\}$ and $\{\hat{\phi}^m_{c,k}\}$. We use mean–squared error for the spectrum terms and the standard VQ codebook commitment loss with stop–gradient $\mathrm{sg}[\cdot]$:

$$\mathcal{L}_{\text{seg}}(x) = \sum_{i=1}^{N} \Big(\|\hat{A}_i - A_i\|_2^2 + \|\hat{\phi}_i - \phi_i\|_2^2 + \|\mathrm{sg}[\ell_2(p_i)] - \ell_2(v_{z_i})\|_2^2 + \|\ell_2(p_i) - \mathrm{sg}[\ell_2(v_{z_i})]\|_2^2 \Big),$$

and the total objective over the dataset $\mathcal{D}$ is

$$\mathcal{L}_T = \sum_{x \in \mathcal{D}} \mathcal{L}_{\text{seg}}(x).$$

As Van Den Oord et al. (2017), we employ the exponential moving average strategy to make the codebook update more stable.

7.3 Experimental Setup

7.3.1 Data and Preprocessing

We train the tokenizer on the Temple University Hospital (TUH) EEG corpus Obeid & Picone (2016). The archive contains clinical EEG from 14,987 subjects with multiple sessions and over 40 montage configurations. Recordings are sampled between 250–1024 Hz, with the majority at 250 Hz.

We select 22 electrodes from the 10–20 system:

$$\{\text{Fp1}, \text{Fp2}, \text{F7}, \text{F3}, \text{Fz}, \text{F4}, \text{F8}, \text{T1}, \text{T3}, \text{C3}, \text{Cz}, \text{C4}, \text{T4}, \text{T2}, \text{T5}, \text{P3}, \text{Pz}, \text{P4}, \text{T6}, \text{O1}, \text{Oz}, \text{O2}\}.$$

Channels with zero or missing signal over a recording are flagged as bad channels. Signals at bad channels are interpolated from neighboring electrodes using distance–weighted averaging with a maximum neighbor radius of 5 cm. All recordings are then re-referenced to the average of the 22 channels.

We then remove power line noise (60 Hz) with a notch filter and apply a band-pass filter (1–48 Hz). All recordings are re-sampled to 200 Hz. A DC-offset correction is applied, and linear trends are removed. Finally, each recording is normalized with a z-transform along the time dimension.

For the tokenizer training, we set the standard input as: 8 chunks, each with a length of 1 second and a 0% overlap. We randomly select a starting point for each EEG recording and then sample 8 contiguous chunks.

7.3.2 Implementation Details and Training Setup

As backbone, the encoder–decoder operates on per–channel patches and feeds a lightweight Transformer over patch embeddings; the decoder mirrors this with a Transformer and two regression heads. The encoder stacks L=12 pre-norm self-attention blocks with H=10 heads, embedding width D=200, and feed-forward width $4D$. Layer normalization with $\varepsilon = 10^{-6}$ is applied both in the block pre-norm and to the query/key projections (qk-norm) before the attention dot-products; activations are GELU, and residual connections are used throughout. After the encoder, token vectors are cosine-normalized and quantized by nearest-neighbor lookup in a codebook of size KxD (2048 x 256) The quantized sequence $\{\hat{v}_{c,k}\}$ is then passed to a decoder that shares the same model width D and applies a small stack of self-attention blocks followed by average pooling over the patch axis. Two linear heads map the decoder output to the spectral targets for each patch, $\hat{A}_{c,k} \in \mathbb{R}^w$ and $\hat{\phi}_{c,k} \in \mathbb{R}^w$.

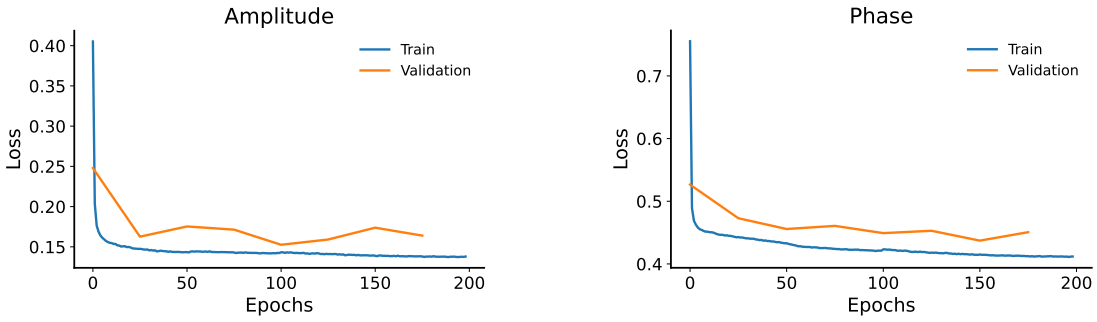
Training is performed on a single NVIDIA RTX 5000 GPU. We use AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.99$, weight decay 1×10^{-5}, batch size 128, and a cosine schedule with warmup learning rate 1×10^{-6}, peak learning rate 5×10^{-5}, and minimum learning rate 1×10^{-5}. 10 warmup epochs are used; total training runs for 100 epochs with checkpoints saved every 10 epochs.

7.4 Pre-Training Results

Figure 7.2 reports the per-epoch reconstruction losses for amplitude and phase on the validation split. The amplitude objective drops in the first few epochs and then stabilizes at a low plateau, suggesting that the model quickly captures the dominant spectral envelope. Phase reconstruction remains more challenging, beginning at a higher level and decreasing more gradually. Nevertheless, the trend is consistently downward throughout training. The difference between the amplitude and phase objectives is expected for low-SNR EEG and is consistent with the observation that spectral shape is easier to recover than precise phase relations.

Additionally, Figure 7.3 illustrates qualitative examples from unseen segments. We present the following: (i) the original time-domain eeg signal, (ii) the original and reconstructed amplitude spectra, and (iii) the original and reconstructed phase. The tokenizer preserves band-limited structure (e.g., alpha/beta peaks) while smoothing high-frequency fluctuations. Phase is less precise, yet remains correlated with the target, particularly at low and mid frequencies.

Across runs, code perplexity increases rapidly during the warm-up phase and then plateaus. The fraction of active codes exceeds 70% after 50 epochs. Dead codes are rare after the first 10–15 epochs, confirming the healthy utilization of the discrete vocabulary. The perplexity value is 1971.5.



(a) Amplitude loss vs. epochs.

(b) Phase loss vs. epochs.

Figure 7.2: Reconstruction losses during training. Amplitude converges quickly to a low plateau; phase decreases more slowly but steadily.

7.5 Brain Tokens as Foundation Model Inputs

The vector-quantized EEG model introduced in the previous sections transforms continuous neural activity into discrete brain tokens, effectively bridging EEG and modern foundation-model architectures. Each EEG segment $\mathbf{x} \in \mathbb{R}^{C\times T}$ is encoded by the VQ-VAE encoder E_q and quantized using a codebook $\mathcal{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$, resulting in a sequence of token indices

$$\mathbf{y} = \{y_1, \ldots, y_N\}, \quad y_t \in [1..K].$$

These tokens serve as compact symbolic representations of neural dynamics, replacing noisy, high-dimensional waveforms with semantically meaningful units that capture dominant spectral–spatial patterns of brain activity.

We connect this discrete tokenizer to an autoregressive GPT model G_θ trained to predict masked tokens in a causal fashion. The GPT receives token embeddings augmented with positional embedding, enabling temporal reasoning across sequences. Given a masked input sequence $\tilde{\mathbf{y}}$, the model

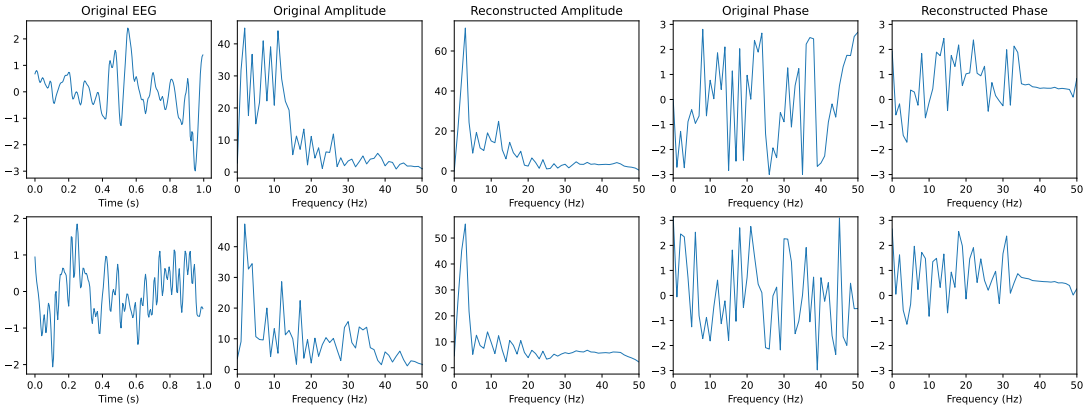


Figure 7.3: Visualization of reconstructed Fourier spectrum on a held–out segment. For the amplitude spectrum, the tokenizer preserves the dominant band-limited structure (e.g., alpha/beta peaks) while smoothing high-frequency noise. While for the phase spectrum, the reconstruction is less precise but remains correlated with the original, particularly at low and mid frequencies.

predicts the missing token $\hat{y}_t = G_\theta(\tilde{\mathbf{y}})$ and minimizes the cross-entropy loss

$$\mathcal{L}_{\text{MTM}} = \mathbb{E}_{t\sim\mathcal{M}}[-\log P_\theta(y_t \mid \tilde{\mathbf{y}}_{<t})]\,,$$

where $\mathcal{M}$ denotes the set of masked positions. This pretraining objective allows the transformer to model multi-scale temporal dependencies and contextual dynamics in EEG signals, analogous to language modeling over discrete vocabularies. An overview of the architecture is shown in Figure 7.4.

7.5.1 Implementation Details

Downstream Dataset. We use the BCI Competition IV Dataset 2a from Graz University of Technology Brunner et al. (2008) for downstream evaluation. Nine subjects performed four motor imagery tasks (left hand, right hand, feet, and tongue), which were recorded using 22 Ag/AgCl electrodes at a sampling rate of 250 Hz. Two sessions were acquired on different days for each subject. Each session contains 72 trials per class (288 trials total). All trials from both sessions are used as training or testing samples, with no overlap between subjects in the splits. The signals are band-pass filtered

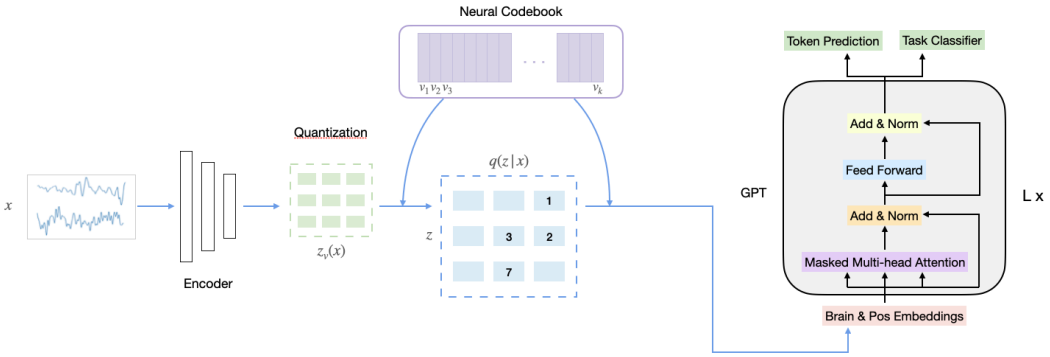


Figure 7.4: The neural tokenizer takes chunks of EEG data as input and generates discrete codes as tokens for the GPT model. The GPT model then predicts the masked brain token or classify the downstream task.

(0.5–100 Hz) and z-score normalized per trial. For each trial, we extracted the interval [2 s, 6 s] relative to the cue, which covered the period during which the imagery task was performed.

The downstream montage differs slightly from the canonical 22-channel configuration used during pre-training. To match the input space, we remap the downstream signals to the canonical layout using a 22×22 linear transformation matrix. This matrix is obtained by solving the forward model and the corresponding inverse mapping for the two sensor configurations. Then, we project the data back to the target montage via the cortical source space. This step produces inputs in the same channel order as pretraining and allows for the direct application of the learned tokenizer.

For the tokenizer inference, we set the standard input as: 4 chunks, each with a length of 1 second and a 0% overlap. We randomly select a starting point for each EEG recording and then sample 4 contiguous chunks within the cue window.

PRETRAINING. We pretrain the brain-token GPT on the TUH EEG corpus. Recordings are segmented into 32 consecutive chunks per training sequence, each chunk spanning 1 second with 50% temporal overlap. The vector-quantized tokenizer is frozen during GPT pretraining, mapping each chunk to a discrete code $y_t \in [1..K]$.

We use a decoder-only transformer G_θ operating over the discrete brain-token vocabulary of size

2048. Tokens are embedded in a 1024-dimensional space. The transformer stacks 6 decoder blocks with masked self-attention and MLP sublayers; layer norm and residual connections follow the standard GPT design. The final linear head projects hidden states to token logits in $\mathbb{R}^K$.

7.5.2 Downstream Task Evaluation

We evaluated the pretrained VQ–EEG tokenizer and GPT model on the BCI–IV–2a motor–imagery benchmark to assess downstream transferability. For each trial, the frozen tokenizer $(E_q, \mathcal{E})$ encodes new EEG signals into tokens, which are then processed by G_θ for semantic decoding and feature extraction.

Finetuning was performed for 10,000 steps using AdamW (lr $= 1 \times 10^{-4}$) and a cross–entropy objective with balanced class weights.

The model achieved a four-class classification accuracy of approximately 35% in a leave-one-subject-out evaluation setting. While this remains below the performance of specialized motor-imagery networks, such as Neuro-GPT Cui et al. (2024) (linear method, 39.8%), it is important to note that our setup does not involve task-specific fine-tuning, unlike Neuro-GPT. Both the tokenizer and the GPT backbone have been trained solely for unsupervised representation learning on heterogeneous clinical EEG datasets and were kept frozen during the downstream evaluation.

Due to the difference in experimental protocol, the comparison is not strictly equivalent. Nevertheless, the obtained accuracy is encouraging. In fact, simple statistical features extracted from the discrete codes and fed into a shallow MLP—without GPT embeddings—yielded only 30–32% accuracy. The additional gain when coupling the same codes with a pretrained GPT highlights the potential of discrete brain tokens as transferable neural interfaces between EEG and foundation-model representations.

At the same time, these results reveal limitations: discrete representations alone cannot fully capture generalizable features across heterogeneous EEG datasets. This emphasizes the need for larger,

more harmonized corpora for cross-dataset training and calibration.

However, discretization introduces several advantages. First, brain tokens drastically reduce data dimensionality and act as denoising abstractions, improving model stability across sessions and subjects. Second, sequences of discrete codes can be processed using the same infrastructure as text. This enables efficient transformer training, perplexity monitoring, and multimodal fusion with linguistic or visual tokens. Finally, the shared token interface supports downstream tasks, such as EEG-to-text decoding, emotion recognition, and motion synthesis, allowing these tasks to reuse pretrained GPT embeddings as a universal representational space for neural signals.

Part V

Conclusions

The only way of discovering the limits of the possible is to venture a little way past them into the impossible.

Arthur C. Clarke

8

Conclusions and Future Directions

8.1 Summary of Contributions

This dissertation focuses on the limitations of traditional EEG-based brain-computer interfaces. Existing systems rely on handcrafted features designed to specific tasks, subjects, and recording sessions. When applied to new users or different contexts, these models suffer catastrophic performance degradation—accuracy drops in cross-subject scenarios, and session-to-session variability introduces comparable instability. Also, traditional EEG systems rely on manually engineered features designed for specific tasks, subjects, and recording conditions. When applied to new contexts, these models suffer catastrophic performance degradation—accuracy drops by 17-19% in cross-subject scenarios, and session-to-session variability introduces comparable instability.

This dissertation demonstrates that learned representations, discovered directly from data through end-to-end training, overcome these limitations by capturing invariant neural patterns while accommodating individual differences. It makes three main contributions:

- Architectures for general EEG representations: it proposes frequency-aware multi-band encoders and subject-conditional transformers for EEG processing. These include frequency-aware multi-band encoders that process band components (delta, theta, alpha, beta) independently while preserving neurophysiological interpretability; subject-conditional layers that explicitly model inter-subject variability through learned embeddings; Transformer Encoders adapted for variable-length EEG sequences with bidirectional temporal feature extraction; and vector-quantized neural tokenization that converts continuous signals into discrete "brain tokens" through learned codebooks, enabling compact representations and lightweight downstream models.

- Semantic decoding from non-invasive neural signals: it presents one of the first end-to-end systems for open-vocabulary EEG-to-text decoding, aligning learned neural representations with

pretrained language models through a two-stage training procedure. Integration of GPT-4 for semantic refinement improved grammatical coherence and factual accuracy. The system was validated with 30 subjects engaged in natural reading and achieved state-of-the-art performance, demonstrating that meaningful linguistic content can be extracted from scalp-recorded brain activity.

- Evaluation methodologies and practical validation: it establishes evaluation frameworks across three distinct paradigms: motor imagery classification, emotion recognition, and semantic decoding. This demonstrates the versatility of learned representations. The introduction of BERTScore to neural decoding evaluation provides a semantic-aware assessment that goes beyond lexical metrics. Comprehensive ablation studies isolated the contribution of each architectural component. All methods were validated across diverse hardware configurations, demonstrating robustness to real-world constraints.

These contributions result in performance improvements: motor imagery classification achieved 66.03% accuracy (compared to 55% with handcrafted features and SVM); emotion recognition with personalized models showed 18% improvement over subject-independent approaches; and semantic decoding reached BLEU-1 of 42.75%, ROUGE-1-F of 33.28%, and BERTScore-F of 53.86%. Discrete tokenization achieved full codebook utilization, demonstrating 32% four-class classification accuracy on the BCI–IV–2a dataset as a downstream task pretrained on heterogeneous clinical EEG.

These results demonstrate that representation learning can extract robust, transferable features from EEG signals, bridging the gap between neural activity and human-interpretable meaning. The following sections synthesize the lessons learned from this work and examine its implications for the future of brain-computer interfaces.

8.2 Learning EEG Representations

This work provides several fundamental insights into learning representations from EEG signals that extend beyond the specific tasks investigated. The most consistent finding across all three experimental paradigms (motor imagery, emotion recognition, and semantic decoding) is that neural networks discovering features through end-to-end training outperform manually engineered pipelines. Traditional approaches based on spectral power, common spatial patterns, or event-related potentials achieved reasonable performance within controlled settings, but learned representations proved consistently superior. The magnitude of improvement varied with task complexity and available data, ranging from 11% gains in motor imagery to 18% in personalized emotion recognition and 3.2% in semantic decoding as measured by BERTScore. The consistency observed across diverse tasks demonstrates that the inductive biases inherent in convolutional and transformer architectures, such as local feature detection, hierarchical composition, and attention mechanisms, align well with the spatiotemporal structure of neural signals.

However, this success came with an important consideration. The 19% accuracy degradation observed when testing subject-independent models on new individuals revealed that inter-subject variability cannot be treated as simple noise that can be averaged away. Anatomical differences in skull thickness, cortical folding patterns, and electrode placement create subject-specific linear transformations between neural sources and scalp measurements. Instead of addressing this variability through extensive data augmentation or domain adaptation techniques alone, we found that modeling subject identity as a latent variable provided a more effective solution. The subject-conditional layers enabled the network to maintain a shared representational space that captures universal neural patterns while accommodating individual differences. While the resulting 1-1.5% performance improvement may seem modest in absolute terms, it represents meaningful progress toward subject-independent systems. More importantly, this improvement demonstrates that treating subject variability as a struc-

tured, learnable phenomenon rather than random noise is essential for deployable BCIs.

These architectural advances address a broader challenge exposed by our transfer learning experiments: dataset heterogeneity. Although pretraining on large clinical EEG corpora, such as TUH, and fine-tuning on task-specific datasets demonstrated that representations can transfer within the EEG modality, the gains were substantially smaller than those routinely observed in computer vision or natural language processing. The fundamental problem is that EEG lacks the standardization that makes transfer learning so effective in other domains. Unlike images, which have universal pixel grids, or text, which has shared vocabularies, EEG recordings vary in electrode montages, sampling rates, reference schemes, and preprocessing pipelines. Two datasets that appear to measure the same phenomenon may be incompatible at the signal level. This is a fundamental barrier to realizing foundation models for neural signals.

8.2.1 The Continuous vs. Discrete Representations Trade-Off

A central architectural decision in neural signal processing is how to represent brain activity: as continuous vectors in high-dimensional space or as discrete tokens from a finite vocabulary. This work explores both paradigms. Rather than identifying a clear winner, we found that the two approaches have complementary strengths, suggesting that different use cases call for different methods.

Table 8.1: Comparison of continuous and discrete EEG representations.

Aspect	Continuous	Discrete
Fidelity	High—preserves subtle patterns	Moderate—information bottleneck
Memory footprint	Large (D-dimensional per segment)	Compact (K codebook indices)
Interpretability	Difficult to analyze directly	Token-level analysis achievable
Downstream tasks	Requires full fine-tuning	Lightweight task-specific heads
Compositionality	Limited	High—tokens combine like words
Best applications	High-stakes medical decoding	Real-time BCIs, edge deployment

This work explored both paradigms and found that they have complementary strengths rather

than one being clearly superior to the other (see Table 8.1). Continuous representations, as demonstrated by our 1,024-dimensional Brain Transformer Encoder embeddings, are ideal for preserving fine-grained information. Examples include medical diagnostics requiring the detection of subtle pathological markers and semantic decoding demanding sensitivity to nuanced distinctions in neural activity. However, they are computationally costly, requiring high-dimensional processing and substantial storage, and offering limited compositionality. On the other hand, vector quantization via discrete codebooks achieves dramatic compression by mapping EEG segments to finite vocabularies, but it introduces an information bottleneck that forces semantically meaningful distinctions. Discrete tokens align naturally with language model architectures and enable lightweight downstream classifiers. This makes them ideal for real-time BCIs and edge deployment scenarios where computational resources are limited.

The key insight is that these approaches can cooperate in hybrid architectures. Continuous encoding captures rich neural information, and vector quantization enables efficient processing and storage. Then, expansion back to continuous space facilitates the final prediction. This flexibility, or the ability to select appropriate representations for each processing stage based on task requirements, enables more valuable than committing exclusively to either paradigm.

8.2.2 Bridging Neural Signals and Semantic Meaning

EEG-to-text experiments demonstrate that extracting semantic content from non-invasive neural recordings is achievable, although it reveals fundamental challenges. Our two-stage approach, which first aligns EEG representations with language model embeddings and then fine-tunes them for generation, proved more effective than end-to-end training. This demonstrates that the gap between raw neural signals and linguistic meaning requires intermediate representations that respect both modalities.

The choice of evaluation metrics profoundly shapes our understanding of progress. Traditional

lexical metrics, such as BLEU, measure word overlap, whereas semantic metrics, such as BERTScore, capture meaning preservation, even when the wording differs. For brain decoding applications, where expressing intent is more important than exact phrasing, semantic evaluation provides a more appropriate assessment. Our ablation studies revealed three distinct challenges: extracting robust features from noisy EEG (7.24% BLEU improvement with better encoders), aligning neural and linguistic spaces (14.25% improvement with language model fine-tuning), and generating coherent text (2.5% improvement with GPT-4 refinement).

The substantial improvement from post-processing refinement demonstrates that current limitations primarily lie in the quality of neural encoding rather than in language generation capability. The gap between our 42.75% BLEU-1 score and the 72.45% upper bound achieved with perfect word identification demonstrates that developing better EEG encoders through larger-scale pretraining, architectural innovations, or multimodal integration represents the most promising direction for future improvement. These findings establish that, although semantic decoding from scalp EEG is feasible, substantial work remains to achieve the necessary robustness and accuracy for practical applications.

8.3 Future Directions and Closing Remarks

This dissertation demonstrates that representation learning is a valid approach for EEG-based brain-computer interfaces. However, it also reveals significant opportunities for improvement. This section outlines promising research directions and concludes with reflections on the path toward practical, generalizable neural decoding systems.

8.3.1 Future Directions

Several research directions can address current limitations and advance EEG-based BCIs toward practical deployment.

Methodological improvements include few-shot learning techniques that enable rapid subject adaptation with minimal calibration data, hierarchical tokenization schemes that capture multi-scale temporal structure, and multi-task pre-training across motor, emotion, and semantic paradigms to discover general-purpose representations. Autoregressive decoding without teacher forcing and improved temporal alignment methods would provide a more realistic evaluation of semantic decoding capabilities.

Scaling toward foundation models requires aggregating diverse EEG datasets for large-scale pre-training with self-supervised objectives such as masked prediction and contrastive learning. Architectural innovations—including neural architecture search tailored to EEG and efficient alternatives to Transformers—may better capture spatiotemporal dependencies. The community needs standardized benchmarks and evaluation protocols to enable systematic comparison across methods. Cross-device learning that achieves montage-invariant representations would enable truly plug-and-play BCIs.

Multimodal integration and applications offer additional opportunities. Combining EEG with fMRI, MEG, eye tracking, and behavioral signals could leverage complementary strengths of different modalities. Advanced applications include bidirectional brain-computer communication with neurofeedback, thought-to-image generation, and mental interfaces for productivity. Real-world deployment requires wearable form factors, continuous monitoring with on-device processing, and privacy-preserving computation.

Ethical considerations must be addressed proactively. Neural data requires robust privacy protections and respect for cognitive liberty. Ensuring equal access requires open-source development and design for different populations. Informed consent frameworks must clearly account for current capabilities and future improvements in decoding technology.

8.3.2 Closing Remarks

This dissertation began with a fundamental question: Can we transition from task-specific EEG systems to more generalizable ones? The answer is yes, through representation learning that discovers invariant features directly from data rather than engineering features for each application.

The evidence is compelling. Learned representations consistently outperform handcrafted features across motor imagery, emotion recognition, and semantic decoding. Subject-conditional architectures accommodate individual differences without sacrificing generalizability. Cross-modal alignment bridges the gap between neural activity and semantic meaning. Neural tokenization enables efficient processing on edge devices. These are not just incremental improvements; they represent a paradigm shift in how we process neural signals.

However, technical achievements mean nothing without real-world impact. The 66% accuracy in motor imagery classification enables reliable wheelchair control in response to user intent. The 42.75% BLEU-1 score in semantic decoding restores communication for individuals with locked-in syndrome. Sub-5 ms inference time enables real-time BCIs on consumer hardware.

For decades, brain-computer interfaces have promised direct neural communication. This dissertation demonstrates that, when combined with large-scale pretraining and modern deep learning architectures, representation learning finally makes that promise achievable. The foundations are established. The path forward is clear. The future of BCIs lies in general-purpose representations that can understand neural activity in all its complexity.

Discovery consists of seeing what everybody has seen and thinking what nobody has thought.

Albert Szent-Gyorgyi

9

Scientific Publications

The following publications were produced during the course of this doctoral research:

Journal Articles

1. **H. Amrani**, D. Micucci, M. Mobilio, P. Napoletano. "Leveraging Dataset Integration and Continual Learning for Human Activity Recognition." *International Journal of Machine Learning and Cybernetics*, pp. 1–22, 2025. Amrani et al. (2025)

2. I.E. Stan, **H. Amrani**, P. Napoletano, D. D'Auria. "Authenticated Robotic Teleoperation with Task Recognition." *IEEE Consumer Electronics Magazine*, 2025. Stan et al. (2026)

3. **H. Amrani**, D. Micucci, P. Napoletano. "Deep Representation Learning for Open Vocabulary Electroencephalography-to-Text Decoding." *IEEE Journal of Biomedical and Health Informatics*, 2024. Amrani et al. (2024b)

Conference Proceedings

1. **H. Amrani**, D. Micucci, P. Napoletano. "Deep Multi-band EEG Learning for Motor Imagery Classification with Dry Electrodes." In *Proceedings of the International Conference on Image Analysis and Processing (ICIAP)*, pp. 339–350, 2025. Amrani et al. (2026)

2. **H. Amrani**, D. Micucci, M. Nalin, P. Napoletano, I. Rizzi. "EEG Acquisition and Motor Imagery Classification for Robotic Control." In *Proceedings of the 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024. Amrani et al. (2024a)

3. A. Fornaro, D. D'Auria, **H. Amrani**, P. Napoletano. "Responsive Teleoperation of a Robotic Arm via Wearable Inertial Sensors." In *Proceedings of the IEEE Gaming, Entertainment, and Media Conference (GEM)*, pp. 1–6, 2024. Fornaro et al. (2024)

4. **H. Amrani**, D. Micucci, M. Nalin, P. Napoletano. "Emotion Personalization with Machine Learning using EEG Signals and Dry Electrodes." In *Proceedings of the IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, 2023. Amrani et al. (2023a)

5. **H. Amrani**, D. Micucci, P. Napoletano. "Unsupervised Deep Learning-based Clustering for Human Activity Recognition." In *Proceedings of the IEEE 12th International Conference on Consumer Electronics (ICCE-Berlin)*, 2022. Amrani et al. (2022)

Part VI

Related Contributions

Progress in science depends on new techniques, new discoveries and new ideas, probably in that order.

Sydney Brenner

10

Supporting Studies

This chapter presents additional works conducted in parallel to the main contributions of this thesis. While not central to the core narrative of representation learning for EEG, these studies were essential for developing the methodological foundations, validating design choices, and exploring alternative approaches that informed the architectures and approaches presented in Parts II, III, and IV. Each section provides a concise summary of the work, key results, and its connection to the broader thesis.

10.1 Unsupervised deep learning-based clustering for human activity recognition

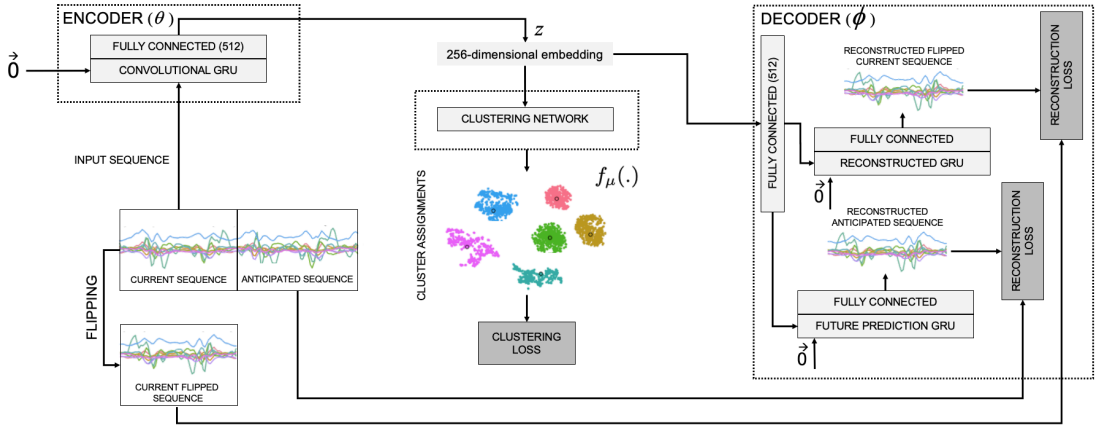


Figure 10.1: Proposed architecture for unsupervised clustering for ADLs. It consists of an autoencoder that tries to reconstruct two outputs from the input signals: the inverse input and the future sequence of the input. Then it is devoted to clustering the dimensionally reduced signals in the learned latent space.

One of the main problems in applying deep learning techniques to recognize activities of daily living (ADLs) based on inertial sensors is the lack of appropriately large labeled datasets to train deep learning-based models. A large amount of data would be available due to the wide spread of mobile devices equipped with inertial sensors that can collect data to recognize human activities. Unfortunately, this data is not labeled. This work proposes DISC (Deep Inertial Sensory Clustering), a DL-based clustering architecture that automatically labels multi-dimensional inertial signals. In particular, the architecture, reported in Figure 10.1, combines a recurrent AutoEncoder and a clustering criterion to predict unlabelled human activity-related signals. The proposed architecture is evaluated on three publicly available HAR datasets and compared with four well-known end-to-end deep clustering approaches. The experiments demonstrate the effectiveness of DISC on both clustering accuracy and normalized mutual information metrics.

10.1.1 Contribution to Thesis

This work validated unsupervised representation learning from time series, directly informing the self-supervised approaches in Chapters 6-7. The clustering-based discretization prefigured vector quantization for brain tokens, while cross-dataset generalization motivated the analysis of EEG variability and subject-conditional architecture design in Chapters 4 and 6.

10.2 Leveraging dataset integration and continual learning for human activity recognition

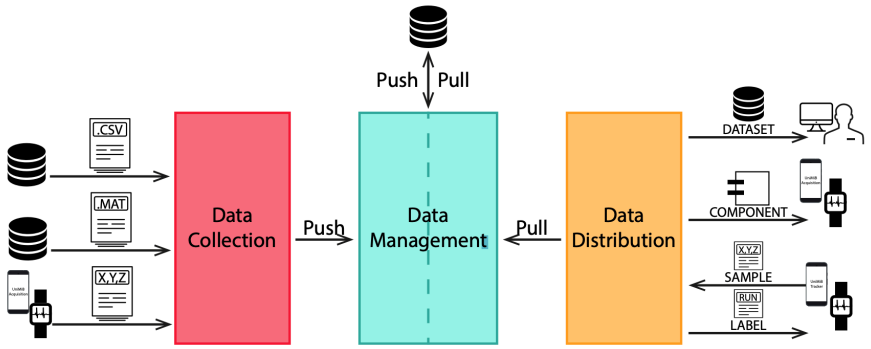


Figure 10.2: CLP is composed of three main components: (i) Data Collection, (ii) Data Management, and (iii) Data Distribution. Data Collection allows the acquisition of a new dataset, while Data Management processes the new dataset to homogenize it and adds it to the existing incrementally built dataset. Finally, the Data Distribution component of CLP enables users and applications to request and receive homogenized sets of labeled signals and custom-trained classifiers from the platform.

Machine learning techniques have proven to be effective in human activity recognition (HAR) from inertial signals. However, they often suffer from intra-class variability and inter-class similarity problems due to strong differences among individuals and in how they perform activities. Recently, data-centric approaches have demonstrated efficacy; however, they require extensive datasets encompassing numerous readings across multiple subjects, incurring significant costs during acquisition campaigns. This study introduces a novel homogenization procedure to address dataset hetero-

geneity in HAR, enabling the integration of diverse datasets into a unified framework. Using eight publicly available HAR datasets, we evaluated the performance of two neural network architectures, a simplified convolutional neural network (S-CNN) and a long short-term memory (LSTM) network. The proposed method reduces the F1-score gap with baseline models from 24.3 to 7.8% on average, reflecting a relative improvement of 16.5%. Additionally, fine-tuning improves model adaptability, achieving a 2.5% accuracy increase for new users. These findings highlight the feasibility of data-centric strategies for robust HAR systems. In particular, the merging procedure, combined with fine-tuning techniques, confirms that diverse data sources and appropriate adaptation methods can yield performance outcomes closely resembling those of the original datasets. Our methodology has been implemented in the continual learning platform (CLP), reported in Figure 10.2, which has been made available to the scientific community to facilitate future research and applications.

10.2.1 Contribution to Thesis

This work established methodologies for handling dataset heterogeneity and enabling cross-dataset generalization, which informed the multi-dataset evaluation strategy throughout this thesis. The homogenization procedure for integrating diverse HAR datasets with different sensor configurations, sampling rates, and activity definitions is similar to the challenges of combining EEG data across recording devices, electrode montages, and experimental paradigms. The fine-tuning strategy for adapting models to new users validated the subject-conditional approach developed in Chapters 5 and 6, in which learned representations balance shared patterns with individual adaptation. The data-centric philosophy that performance improvements stem from training on diverse data rather than architectural complexity motivated the emphasis on learning generalizable representations from large-scale, heterogeneous corpora instead of engineering specialized models for each BCI paradigm.

10.3 Responsive teleoperation of a robotic arm via wearable inertial sensors

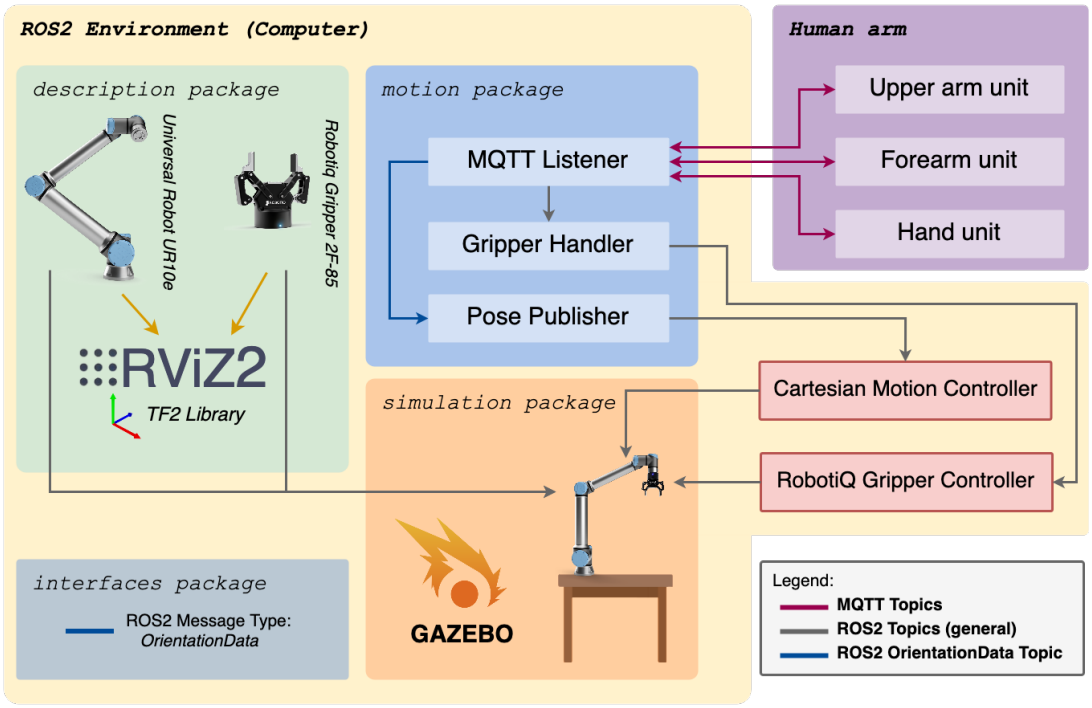


Figure 10.3: Overall architecture of the proposed system. The components highlighted in blue and violet are responsible for motion tracking, while the others are focused on the telecontrol of a robotic arm and its simulation.

Human-in-the-loop teleoperation stands as a primary method for controlling robotic arms, offering significant practical and research applications across various fields such as industry, medicine, and home automation. This work proposes a responsive system for the remote control of a robotic arm utilizing wearable inertial sensors. An overview of the approach is presented in Figure 10.3. The system relies on three inertial units connected to low-cost wireless microcontrollers, which are placed on the user's arm. A server-side application, built on Robotic Operating System (ROS), orchestrates the inter-action between these units and the robotic arm. The robotic arm's functionality is demonstrated within a simulated environment using Gazebo (an open-source 2D/3D robotics simulator).

To prove the efficacy of our proposed system, we conducted an experiment involving 16 users and evaluated their interaction using both objective measures (such as manipulation task execution time) and subjective metrics (including user satisfaction). Our experimentation reveals that users can teleoperate for longer times compared to real manipulation scenarios initially, but can gradually reduce these times as their experience increases. Furthermore, the majority of users rated the interaction as realistic, responsive, and usable.

10.3.1 Contribution to Thesis

The experimental methodology that combined objective performance metrics with subjective usability assessments shaped the evaluation protocols throughout the thesis. Observing users' adaptation during teleoperation was critical because it validated the importance of subject-specific learning, which was addressed in Chapter 4. This validation motivated the subject-conditional architectures in Parts III and IV, which accommodate individual differences.

10.4 Authenticated Robotic Teleoperation with Task Recognition

Human-in-the-loop teleoperation of robotic arms holds considerable promise for consumer electronics, particularly in immersive and interactive applications such as healthcare, gaming, augmented reality, virtual reality, smart home systems, and telemedicine. These applications demand robust user authentication to safeguard against unauthorized access, ensuring secure, user-centered control of teleoperated systems in various consumer environments. However, existing systems often lack seamless integration of security measures, such as biometric authentication, while maintaining affordability and responsiveness. This work proposes a low-cost teleoperation system using wearable inertial measurement units (IMUs) to securely and responsively control a robotic arm. We employ a machine-learning approach for authentication using logistic regression on time series data from the IMUs

Table 10.1: Models, hyperparameters, and cross-validation macro F1-scores. The best hyperparameters for task identification are italicized, while those for user identification are underlined. The scores of the best-performing models are in bold.

Model	Hyperparameters	F1-Score Task Id. (%)	F1-Score User Id. (%)
RF	n_estimators: 25, 50, 75, 100, 125, 150, 175, *200* max_depth: null, 1, 2, 3, 4, 5, 6, 7, 8, *9*, 10 criterion: *gini*, entropy, log_loss	**75.60**	63.15
SVM	C: 0.1, *0.2*, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 kernel: *linear*, rbf, sigmoid	60.06	67.91
KNN	n_neighbors: 1, 3, 5, 7, *9*, 11, 13, 15 weights: *uniform*, distance	64.38	61.02
DT	max_depth: null, 1, 2, *3*, 4, 5, 6, 7, 8, 9, 10 criterion: *gini*, entropy, log_loss	69.09	36.02
NB	var_smoothing: 1e-11, 1e-10, 1e-9, 1e-8, *1e-7*	69.86	43.77
LR	C: 0.1, 0.2, 0.3, 0.4, 0.5, *0.6*, 0.7, 0.8, 0.9, 1.0	68.70	**74.02**

during handling tasks. Experiments with 16 operators performing three handling tasks demonstrate that random forest outperforms other classifiers in task identification, achieving a macro F1-score of 75.60%. In contrast, logistic regression performs best in user identification and authentication tasks. Our system achieves an average Equal Error Rate of approximately 8.89% in user authentication using logistic regression. Results are reported in Table 10.1. The proposed system's low-cost, IMU-based design, adaptable to various end-effectors, aligns with consumer demands for affordable, intuitive, and secure teleoperation setups. This work highlights the potential for biometric-based teleoperation to advance consumer technology applications in healthcare, Internet of Things, and immersive environments, ensuring personalized and secure user experiences.

10.4.1 Contribution to Thesis

This work reframed individual differences in sensor signals as discriminative features rather than as source of variability. This approach achieved an Equal Error Rate of 8.89% for user authentication. This insight motivated the development of subject-conditional architectures, in Chapters 5 and 6, that

explicitly model individual patterns as informative, rather than suppressing them. The task recognition methodology in Chapter 3 achieved a 75.60% F1-score and validated machine learning comparison approaches that transferred to motor imagery classification .

10.5 Physics-Based and Physiological Human Motion Diffusion

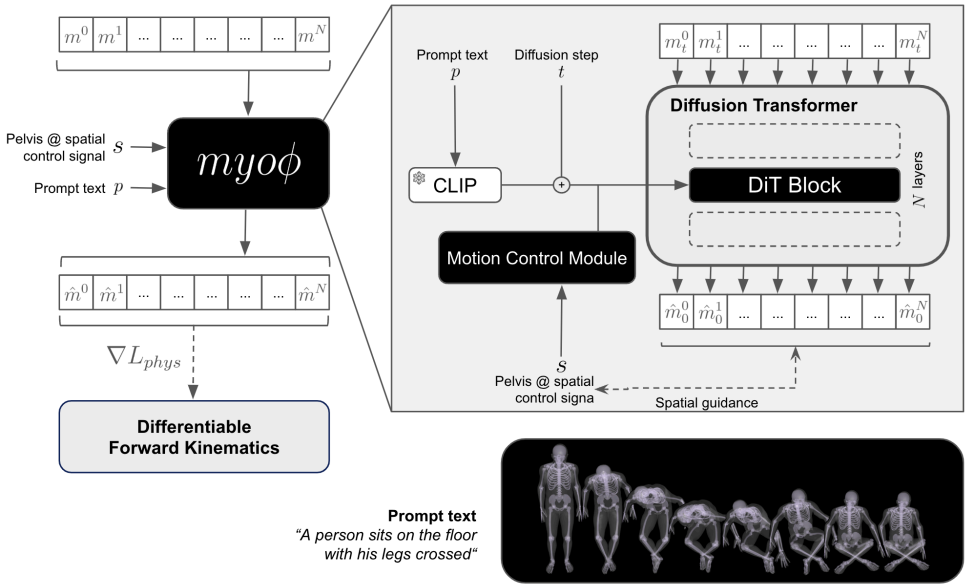


Figure 10.4: Architecture of $myo\Phi$ for physical and physiologically human motion generation. A textual encoder (CLIP) processes language inputs to obtain semantic embeddings, while an optional motion control module incorporates spatial signals (e.g., pelvis trajectories, IMU data). Both signals condition a diffusion transformer that progressively denoises motion states in the parameter space μ_{myo}. A differentiable forward kinematics component enforces realistic joint ranges and contact constraints throughout training, ensuring each refinement step adheres to physical and physiological principles.

Current approaches for text-to-motion generation are capable of generating diverse human behaviors with geometric consistency and produce visually coherent motions, but they use motion representations that neglect the physical and physiological (PnP) consistency that is fundamental to realistic human motion. Addressing this limitation has significant potential for applications in domains

requiring high motion fidelity, such as biomechanics, ergonomics, health, and safety. This work introduces μ_{myo}, a novel parameterization that, in contrast to commonly used motion representations, inherently enforces physiological consistency. It also proposes *myo*Φ, a diffusion model that generates motions leveraging μ_{myo} together with a physiologically accurate body model, the *MyoSkeleton*. The architecture is shown in Figure 10.4. By tightly coupling the motion representation with the body model and by subsequently generating motions in this space, we embed physical and physiological information directly into *myo*Φ, eliminating the need for external constraints such as geometric training objectives or the rollout of control policies in physics simulators—reducing both training time and the number of diffusion steps required during inference. As a result, the proposed model achieves state-of-the-art results, with a Fréchet Inception Distance of 0.170, a 10.5% lower trajectory error, a 36% reduction in ground penetration, and a 41% improvement in the physiological score—all while generating 200 frames in under two seconds.

10.5.1 Contribution to Thesis

This work explored controlled generation through diffusion models on learned representations, establishing principles that transferred to neural signal processing. The key insight, that representation design determines generalization and controllability, informed Parts III-IV. Both human motion and neural activity exhibit hierarchical organization, temporal dynamics, and subject-specific characteristics that must be captured while maintaining generalizability. The diffusion framework validated that generative models produce constrained outputs in structured latent spaces, which motivated reconstruction objectives for brain tokens. Effective representation learning requires aligning learned spaces with domain structure to enable controlled generation and generalization as emergent properties.

References

Abdi, H. & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433–459.

Al-Fahoum, A. S. & Al-Fraihat, A. A. (2014). Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains. International Scholarly Research Notices, 2014(1), 730218.

Ali, M., Mosa, A. H., Al Machot, F., & Kyamakya, K. (2016). Eeg-based emotion recognition approach for e-healthcare applications. In 2016 eighth international conference on ubiquitous and future networks (ICUFN) (pp. 946–950).: IEEE.

Alimardani, M. & Hiraki, K. (2020). Passive brain-computer interfaces for enhanced human-robot interaction. Frontiers in Robotics and AI, 7, 125.

Amrani, H., Micucci, D., Mobilio, M., & Napoletano, P. (2025). Leveraging dataset integration and continual learning for human activity recognition. International Journal of Machine Learning and Cybernetics, (pp. 1–22).

Amrani, H., Micucci, D., Nalin, M., & Napoletano, P. (2023a). Emotion personalization with machine learning using eeg signals and dry electrodes. In 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) (pp. 132–137).

Amrani, H., Micucci, D., Nalin, M., Napoletano, P., & Rizzi, I. (2024a). Eeg acquisition and motor imagery classification for robotic control. In 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 1–4).

Amrani, H., Micucci, D., & Napoletano, P. (2021). Personalized models in human activity recognition using deep learning. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 9682–9688).

Amrani, H., Micucci, D., & Napoletano, P. (2022). Unsupervised deep learning-based clustering for human activity recognition. In 2022 IEEE 12th International Conference on Consumer Electronics (ICCE-Berlin) (pp. 1–6).

Amrani, H., Micucci, D., & Napoletano, P. (2023b). Deep representation learning for open vocabulary electroencephalography-to-text decoding. arXiv preprint arXiv:2312.09430.

Amrani, H., Micucci, D., & Napoletano, P. (2024b). Deep representation learning for open vocabulary electroencephalography-to-text decoding. IEEE Journal of Biomedical and Health Informatics, (pp. 1–12).

Amrani, H., Micucci, D., & Napoletano, P. (2026). Deep multi-band eeg learning for motor imagery classification with dry electrodes. In E. Rodolà, F. Galasso, & I. Masi (Eds.), Image Analysis and Processing – ICIAP 2025 (pp. 339–350). Cham: Springer Nature Switzerland.

Ang, K. K., Chin, Z. Y., Zhang, H., & Guan, C. (2008). Filter bank common spatial pattern (fbcsp) in brain-computer interface. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 2390–2397).: IEEE.

Angrisani, L., Arpaia, P., Donnarumma, F., Esposito, A., Frosolone, M., Improta, G., Moccaldi, N., Natalizio, A., & Parvis, M. (2020). Instrumentation for motor imagery-based brain computer interfaces relying on dry electrodes: a functional analysis. In 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) (pp. 1–6).: IEEE.

Apicella, A., Arpaia, P., Cataldo, A., D'Errico, G., Marocco, D., Mastrati, G., Moccaldi, N., Pollastro, A., Ricciardi, B., & Vallefuoco, E. (2022). Reproducible assessment of valence and arousal based on an eeg wearable device. In 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) (pp. 661–666).: IEEE.

Aristimunha, B., Carrara, I., Guetschel, P., Sedlar, S., Rodrigues, P., Sosulski, J., Narayanan, D., Bjareholt, E., Barthelemy, Q., Schirrmeister, R. T., Kobler, R., Kalunga, E., Darmet, L., Gregoire, C., Abdul Hussain, A., Gatti, R., Goncharenko, V., Thielen, J., Moreau, T., Roy, Y., Jayaram, V., Barachant, A., & Chevallier, S. (2025). Mother of all bci benchmarks.

Arpaia, P., Coyle, D., Donnarumma, F., Esposito, A., Natalizio, A., Parvis, M., Pesola, M., & Vallefuoco, E. (2022). Multimodal feedback in assisting a wearable brain-computer interface based on motor imagery. In 2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) (pp. 691–696).

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

Balakrishnama, S. & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing, 18(1998), 1–8.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1798–1828.

Binnie, C. & Prior, P. (1994). Electroencephalography. Journal of Neurology, Neurosurgery & Psychiatry, 57(11), 1308–1319.

Bos, D. O. et al. (2006). Eeg-based emotion recognition. The influence of visual and auditory stimuli, 56(3), 1–17.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Current Biology, 28(5), 803–809.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877–1901.

Brunner, C., Leeb, R., Müller-Putz, G., Schlögl, A., & Pfurtscheller, G. (2008). Bci competition 2008–graz data set a. Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology, 16(1-6), 34.

Casso, M.-I., Jeunet, C., & Roy, R. N. (2021). Heading for motor imagery brain-computer interfaces (mi-bcis) usable out-of-the-lab: impact of dry electrode setup on classification accuracy. In 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER) (pp. 690–693).: IEEE.

Caucheteux, C. & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. Communications biology, 5(1), 134.

Cecotti, H. (2011). Spelling with non-invasive brain–computer interfaces–current and future trends. Journal of Physiology-Paris, 105(1-3), 106–114.

Chaudhary, U., Birbaumer, N., & Ramos-Murguialday, A. (2016). Brain–computer interfaces for communication and rehabilitation. Nature Reviews Neurology, 12(9), 513–525.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597–1607).: PmLR.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

Cincotti, F., Mattia, D., Aloise, F., Bufalari, S., Schalk, G., Oriolo, G., Cherubini, A., Marciani, M. G., & Babiloni, F. (2008). Non-invasive brain–computer interface system: towards its application as assistive technology. Brain research bulletin, 75(6), 796–803.

Cohen, M. X. (2014). Analyzing neural time series data: theory and practice. Cambridge, MA: MIT press.

Comon, P. (1994). Independent component analysis, a new concept? Signal processing, 36(3), 287–314.

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (eeg) classification tasks: a review. Journal of neural engineering, 16(3), 031001.

Cui, W., Jeong, W., Thölke, P., Medani, T., Jerbi, K., Joshi, A. A., & Leahy, R. M. (2024). Neurogpt: Towards a foundation model for eeg. In 2024 IEEE International Symposium on Biomedical Imaging (ISBI) (pp. 1–5).: IEEE.

Dash, D., Ferrari, P., & Wang, J. (2020). Decoding imagined and spoken phrases from non-invasive neural (meg) signals. Frontiers in neuroscience, 14, 290.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2023a). Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence, 5(10), 1097–1107.

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2023b). Decoding speech perception from non-invasive brain recordings. Nature Machine Intelligence, (pp. 1–11).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171–4186).

Ding, W., Nakai, K., & Gong, H. (2022). Protein design via deep learning. Briefings in bioinformatics, 23.

Ding, Y., Robinson, N., Zeng, Q., Chen, D., Wai, A. A. P., Lee, T.-S., & Guan, C. (2020). Tsception: a deep learning framework for emotion detection using eeg. In 2020 international joint conference on neural networks (IJCNN) (pp. 1–7).: IEEE.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Duan, Y., Zhou, J., Wang, Z., Wang, Y.-K., & Lin, C.-T. (2023). Dewave: Discrete eeg waves encoding for brain dynamics to text translation. arXiv preprint arXiv:2309.14030.

Ekman, P. et al. (1999). Basic emotions. Handbook of cognition and emotion, 98(45-60), 16.

Fahimi, F., Zhang, Z., Goh, W. B., Lee, T.-S., Ang, K. K., & Guan, C. (2019). Inter-subject transfer learning with an end-to-end deep convolutional neural network for eeg-based bci. Journal of neural engineering, 16(2), 026007.

Farwell, L. A. & Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and clinical Neurophysiology, 70(6), 510–523.

Ferree, T. C., Clay, M., & Tucker, D. M. (2001). The spatial resolution of scalp eeg. Neurocomputing, 38, 1209–1216.

Fornaro, A., D'Auria, D., Amrani., H., & Napoletano, P. (2024). Responsive teleoperation of a robotic arm via wearable inertial sensors. In 2024 IEEE Gaming, Entertainment, and Media Conference (GEM) (pp. 1–6).

Gao, Z., Dang, W., Wang, X., Hong, X., Hou, L., Ma, K., & Perc, M. (2021). Complex networks and deep learning for eeg signal analysis. Cognitive Neurodynamics, 15(3), 369–388.

Gauthier, J. & Ivanova, A. (2018). Does the brain represent words? an evaluation of brain decoding studies of language understanding. arXiv preprint arXiv:1806.00591.

Giraud-Carrier, C. (2000). A note on the utility of incremental learning. Ai Communications, 13(4), 215–223.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. Communications of the ACM, 63(11), 139–144.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). Meg and eeg data analysis with mne-python. Frontiers in neuroscience, (pp. 267).

Guetschel, P., Ahmadi, S., & Tangermann, M. (2024). Review of deep representation learning techniques for brain–computer interfaces. Journal of Neural Engineering, 21(6), 061002.

Hairston, W. D., Whitaker, K. W., Ries, A. J., Vettel, J. M., Bradford, J. C., Kerick, S. E., & McDowell, K. (2014). Usability of four commercially-oriented eeg systems. Journal of neural engineering, 11(4), 046018.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

Hendrycks, D. & Gimpel, K. (2016). Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.

Herwig, U., Satrapi, P., & Schönfeldt-Lecuona, C. (2003). Using the international 10-20 eeg system for positioning of transcranial magnetic stimulation. Brain topography, 16, 95–99.

Hinton, G. E. (1984). Distributed representations.

Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011). Transforming auto-encoders. In International conference on artificial neural networks (pp. 44–51).: Springer.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527–1554.

Hoffer, E. & Ailon, N. (2015). Deep metric learning using triplet network. In International workshop on similarity-based pattern recognition (pp. 84–92).: Springer.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. Scientific data, 5(1), 1–13.

Hollenstein, N., Troendle, M., Zhang, C., & Langer, N. (2019). Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. arXiv preprint arXiv:1912.00903.

Hossain, K. M., Islam, M. A., Hossain, S., Nijholt, A., & Ahad, M. A. R. (2023). Status of deep learning for eeg-based brain–computer interface applications. Frontiers in computational neuroscience, 16, 1006763.

Hu, L. & Zhang, Z. (2019). Eeg signal processing and feature extraction. International Journal for Modern Trends in Science and Technology.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. Nature, 532(7600), 453–458.

Izenman, A. J. (2012). Introduction to manifold learning. Wiley Interdisciplinary Reviews: Computational Statistics, 4(5), 439–446.

Jiang, W.-B., Zhao, L.-M., & Lu, B.-L. (2024). Large brain model for learning generic representations with tremendous eeg data in bci. arXiv preprint arXiv:2405.18765.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. Nature, 452(7185), 352–355.

Kim, B. H. & Jo, S. (2018). Deep physiological affect network for the recognition of human emotions. IEEE Transactions on Affective Computing, 11(2), 230–243.

King, J.-R. & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. Trends in cognitive sciences, 18(4), 203–210.

Kingma, D. P. & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Klem, G. H. (1999). The ten-twenty electrode system of the international federation. the international federation of clinical neurophysiology. Electroencephalogr. Clin. Neurophysiol. Suppl., 52, 3–6.

Kostas, D., Aroca-Ouellette, S., & Rudzicz, F. (2021). Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data. Frontiers in Human Neuroscience, 15, 653659.

Kutas, M. & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). Annual review of psychology, 62, 621–647.

Lang, P. J., Bradley, M. M., Cuthbert, B. N., et al. (2005). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. NIMH, Center for the Study of Emotion & Attention Gainesville, FL.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. Journal of neural engineering, 15(5), 056013.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Li, D., Xu, J., Wang, J., Fang, X., & Ji, Y. (2020). A multi-scale fusion convolutional neural network based on attention mechanism for the visualization analysis of eeg signals decoding. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(12), 2615–2626.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74–81).

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in neural information processing systems, 36, 34892–34916.

Lopez-Gordo, M. A., Sanchez-Morillo, D., & Valle, F. P. (2014). Dry eeg electrodes. Sensors, 14(7), 12847–12870.

Lu, H., Eng, H.-L., Guan, C., Plataniotis, K. N., & Venetsanopoulos, A. N. (2010). Regularized common spatial pattern with aggregation for eeg classification in small-sample setting. IEEE transactions on Biomedical Engineering, 57(12), 2936–2946.

Makeig, S., Bell, A., Jung, T.-P., & Sejnowski, T. J. (1995). Independent component analysis of electroencephalographic data. Advances in neural information processing systems, 8.

Mathewson, K. E., Harrison, T. J., & Kizuk, S. A. (2017). High and dry? comparing active dry eeg electrodes to active and passive wet electrodes. Psychophysiology, 54(1), 74–82.

Merk, T., Salehi, S., Koehler, R. M., Cui, Q., Olaru, M., Hahn, A., Provenza, N. R., Little, S., Abbasi-Asl, R., Starr, P. A., et al. (2025). Pre-trained transformer-models using chronic invasive electrophysiology for symptom decoding without patient-individual training. arXiv preprint arXiv:2508.10160.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. Science, 320(5880), 1191–1195.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., et al. (2021). Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine, 385(3), 217–227.

Moyer, J. T., Gnatkovsky, V., Ono, T., Otáhal, J., Wagenaar, J., Stacey, W. C., Noebels, J., Ikeda, A., Staley, K., de Curtis, M., et al. (2017). Standards for data acquisition and software-based analysis of in vivo electroencephalography recordings from animals. a task 1-wg 5 report of the aes/ilae translational task force of the ilae. Epilepsia, 58, 53–67.

Nicolas-Alonso, L. F. & Gomez-Gil, J. (2012). Brain computer interfaces, a review. sensors, 12(2), 1211–1279.

Noah, S., Powell, T., Khodayari, N., Olivan, D., Ding, M., & Mangun, G. R. (2020). Neural mechanisms of attentional control for objects: decoding eeg alpha when anticipating faces, scenes, and tools. Journal of Neuroscience, 40(25), 4913–4924.

Obeid, I. & Picone, J. (2016). The temple university hospital eeg data corpus. Frontiers in neuroscience, 10, 196.

OpenAI (2023). Gpt-4 technical report. ArXiv, abs/2303.08774.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311–318).

Peksa, J. & Mamchur, D. (2023). State-of-the-art on brain-computer interface technology. Sensors, 23(13), 6001.

Peng, W. (2019). Eeg preprocessing and denoising. EEG Signal Processing and Feature Extraction, (pp. 71–87).

Peng, Z., Dong, L., Bao, H., Ye, Q., & Wei, F. (2022). Beit v2: Masked image modeling with vector-quantized visual tokenizers. arXiv preprint arXiv:2208.06366.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. Nature communications, 9(1), 963.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763).: PMLR.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research, 21(1), 5485–5551.

Russell, J. A. (1980). A circumplex model of affect. Journal of personality and social psychology, 39(6), 1161.

Saha, S. & Baumert, M. (2020). Intra-and inter-subject variability in eeg-based sensorimotor brain computer interface: a review. Frontiers in computational neuroscience, 13, 87.

Sakhavi, S., Guan, C., & Yan, S. (2018). Learning temporal information for brain-computer interface using convolutional neural networks. IEEE transactions on neural networks and learning systems, 29(11), 5619–5629.

Scheer, H. J., Sander, T., & Trahms, L. (2005). The influence of amplifier, interface and biological noise on signal quality in high-resolution eeg recordings. Physiological measurement, 27(2), 109.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for eeg decoding and visualization. Human brain mapping, 38(11), 5391–5420.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv, (pp. 407007).

Sen, D., Mishra, B. B., & Pattnaik, P. K. (2023). A review of the filtering techniques used in eeg signal processing. In 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 270–277).: IEEE.

Shoeibi, A., Ghassemi, N., Alizadehsani, R., Rouhani, M., Hosseini-Nejad, H., Khosravi, A., Panahiazar, M., & Nahavandi, S. (2021). A comprehensive comparison of handcrafted features and convolutional autoencoders for epileptic seizures detection in eeg signals. Expert Systems with Applications, 163, 113788.

Simmatis, L., Russo, E. E., Geraci, J., Harmsen, I. E., & Samuel, N. (2023). Technical and clinical considerations for electroencephalography-based biomarkers for major depressive disorder. Npj Mental Health Research, 2(1), 18.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631–1642).

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929–1958.

Stan, I. E., Amrani, H., Napoletano, P., & D'Auria, D. (2026). Authenticated robotic teleoperation with task recognition. IEEE Consumer Electronics Magazine, 15(1), 33–40.

Stancin, I., Cifrek, M., & Jovic, A. (2021). A review of eeg signal features and their application in driver drowsiness detection systems. Sensors, 21(11), 3786.

Sugden, R. J., Pham-Kim-Nghiem-Phu, V.-L. L., Campbell, I., Leon, A., & Diamandis, P. (2023). Remote collection of electrophysiological data with brain wearables: opportunities and challenges. Bioelectronic Medicine, 9(1), 12.

Takagi, Y. & Nishimoto, S. (2023). High-resolution image reconstruction with latent diffusion models from human brain activity. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (pp. 14453–14463).

Tang, J., LeBel, A., Jain, S., & Huth, A. G. (2023). Semantic reconstruction of continuous language from non-invasive brain recordings. Nature Neuroscience, (pp. 1–9).

Tortora, S., Ghidoni, S., Chisari, C., Micera, S., & Artoni, F. (2020). Deep learning-based bci for gait decoding from eeg with lstm recurrent neural network. Journal of neural engineering, 17(4), 046011.

Vaidya, A. R., Jain, S., & Huth, A. G. (2022). Self-supervised models of audio effectively explain human cortical responses to speech. arXiv preprint arXiv:2205.14252.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. Advances in neural information processing systems, 30.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, C., Subramaniam, V., Yaari, A. U., Kreiman, G., Katz, B., Cases, I., & Barbu, A. (2023). Brainbert: Self-supervised representation learning for intracranial recordings. arXiv preprint arXiv:2302.14367.

Wang, R., Wang, J., Yu, H., Wei, X., Yang, C., & Deng, B. (2015). Power spectral density and coherence analysis of alzheimer's eeg. Cognitive neurodynamics, 9, 291–304.

Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., XiXuan, S., et al. (2024a). Cogvlm: Visual expert for pretrained language models. Advances in Neural Information Processing Systems, 37, 121475–121499.

Wang, Y., Turnbull, A., Xiang, T., Xu, Y., Zhou, S., Masoud, A., Azizi, S., Lin, F. V., & Adeli, E. (2024b). Decoding visual experience and mapping semantics through whole-brain analysis using fmri foundation models. arXiv preprint arXiv:2411.07121.

Wang, Z. & Ji, H. (2022). Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36 (pp. 5350–5358).

Yadav, H. & Maini, S. (2023). Electroencephalogram based brain-computer interface: Applications, challenges, and opportunities. Multimedia Tools and Applications, 82(30), 47003–47047.

Yang, C., Westover, M., & Sun, J. (2023). Biot: Biosignal transformer for cross-data learning in the wild. Advances in Neural Information Processing Systems, 36, 78240–78260.

Yi, K., Wang, Y., Ren, K., & Li, D. (2023). Learning topology-agnostic eeg representations with geometry-aware modeling. Advances in Neural Information Processing Systems, 36, 53875–53891.

Zhang, D., Yao, L., Zhang, X., Wang, S., Chen, W., Boots, R., & Benatallah, B. (2018). Cascade and parallel convolutional recurrent neural networks on eeg-based intention recognition for brain computer interface. In Proceedings of the aaai conference on artificial intelligence, volume 32.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019a). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Zhang, Y. (2025). The evolution of neural signal decoding techniques in brain-computer interfaces from traditional methods to deep learning. Theoretical and Natural Science, 133, 216–223.

Zhang, Y., Zhang, X., Sun, H., Fan, Z., & Zhong, X. (2019b). Portable brain-computer interface based on novel convolutional neural network. Computers in biology and medicine, 107, 248–256.

Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592.

Zhuang, M., Wu, Q., Wan, F., & Hu, Y. (2020). State-of-the-art non-invasive brain–computer interface for neural rehabilitation: A review. Journal of Neurorestoratology, 8(1), 12–25.