POLITECNICO DI TORINO

MASTER's Degree in DATA SCIENCE ENGINEERING

MASTER's Degree Thesis

How feature richness shapes EEG decoding across BCI paradigms

Supervisors

Prof. Luca MESIN

Dr. Hossein AHMADI

Candidate

Mohammad Javad ASGARI

December 2025

Abstract

This thesis empirically evaluates how Electroencephalography (EEG) representational form shapes decoding performance across three canonical brain-computer interface (BCI) paradigms: motor imagery (MI), Event-Related Potential (ERP) and Steady-State Visually Evoked Potential (SSVEP). It systematically compares four EEG-native representations Raw time-series, Power Spectral Density (PSD), Spectrograms, Phase-Locking Value (PLV), and Approximate Entropy (ApEn)—paired with four models Linear Discriminant Analysis(LDA), Logistic Regression (LR), Convolutional Neural Network (CNN), Long Short-Term Memory Network (LSTM) under a rigorous, leakage-free evaluation protocol using public datasets (BNCI2014_001, BNCI2014_009, Nakanishi2015).

Using fixed computational budgets and subject-aware splitting, results reveal a clear performance hierarchy. PSD and Raw dominate core metrics (Balanced Accuracy(BA), Macro-F1, Cohen's κ, AUROC, AUPRC) when combined with deep architectures, achieving near-ceiling performance for ERP and substantial gains for MI and SSVEP. Deep learning (DL)amplifies these advantages but cannot rescue weak representations: ApEn consistently underperforms across all conditions, while PLV provides only modest MI utility. CNNs demonstrate superior robustness versus LSTMs, particularly for frequency-tagged signals.

These findings support a conditional Representation–Richness Principle(RRP): representational benefits materialize only when preserved structure aligns with paradigm-specific neurophysiology and model inductive biases. The study provides reproducible benchmarks and practical guidelines—classical baselines remain competitive for resource-constrained applications, while DL on Raw offers optimal accuracy–cost trade-offs for high-performance BCIs.

Keywords: EEG; ML ; DL; PSD; spectrogram; PLV; ApEn; ERP; MI; SSVEP; calibration.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisors, Professor Luca Mesin and Dr. Hossein Ahmadi, whose invaluable guidance, support, and encouragement have been fundamental to the completion of this thesis. Their insightful advice and continuous mentorship have greatly enriched my academic journey. My heartfelt thanks go to my beloved parents, whose unconditional love, prayers, and constant support have always been my source of strength and inspiration. I am also sincerely grateful to my dear friend, Shahin Saeedpour, for his kindness, encouragement, and unwavering companionship throughout this journey.

Table of Contents

List of Figures

List of Tables

Acronyms

EEG	Electroencephalography.
BCI	Brain-Computer Interface.
MI	Motor Imagery.
ERP	Event-Related Potential.
SSVEP	Steady-State Visually Evoked Potential.
fMRI	Functional Magnetic Resonance Imaging.
PSD	Power Spectral Density.
PLV	Phase Locking Value.
ApEn	Approximate Entropy.
SNR	Signal-to-Noise Ratio.
SPD	Symmetric Positive-Definite.
CNN	Convolutional Neural Network.
LSTM	Long Short-Term Memory Network.
LDA	Linear Discriminant Analysis.
LR	Logistic Regression.
DL	Deep Learning.
ML	Machine Learning.
MOABB	Mother of All BCI Benchmarks.
SST	Single Source of Truth.

BA	Balanced Accuracy.
F1	F1-score.
Macro-F1	Macro-averaged F1-score.
κ	Cohen's Kappa.
AUROC	Area Under the Receiver Operating Characteristic Curve.
AUROC_macro	Macro-averaged AUROC.
AUPRC	Area Under the Precision-Recall Curve.
AUPRC_macro	Macro-averaged AUPRC.
RRP	Representation Richness Principle.
NPZ	NumPy compressed binary format.

Chapter 1

Introduction

The brain, the command center of the nervous system, is a remarkably intricate and dynamic organ responsible for all conscious thought, emotion, behavior, and fundamental bodily functions. It is a complex biological entity, composed of billions of interconnected neurons, glial cells, and supporting tissues, all meticulously organized into specialized regions and networks. This highly organized structure allows for the continuous reception, processing, and transmission of information that underpins our perception of the world, our ability to learn and remember, and our capacity to interact with our environment. Ultimately, the brain serves as the foundation of our individuality, shaping our experiences and defining what it means to be human. The various regions of the brain are interconnected through a complex network of neural pathways, utilizing electrical and chemical signals to communicate. This connectivity enables the integration of information across different brain areas, which is essential for the coordination and execution of a wide range of cognitive, sensory, and motor functions.

Accessing the intricate workings of the brain is a significant challenge, yet crucial for understanding its functions in both health and disease. Various methods have been developed to probe its structure and activity, each offering unique insights at different levels of resolution. Invasive and non-invasive are terms used to categorize techniques for studying or interacting with the brain based on whether they require physically breaching the skull and directly accessing neural tissue.

- **Invasive techniques.** Invasive techniques involve physically accessing the brain tissue, typically through surgical procedures. These methods allow for direct interaction with neurons and neural circuits.

- **Non-invasive techniques.** Those that do not involve any surgical procedures or direct physical penetration into the brain.These methods typically record or stimulate brain activity from the scalp or outside the body.

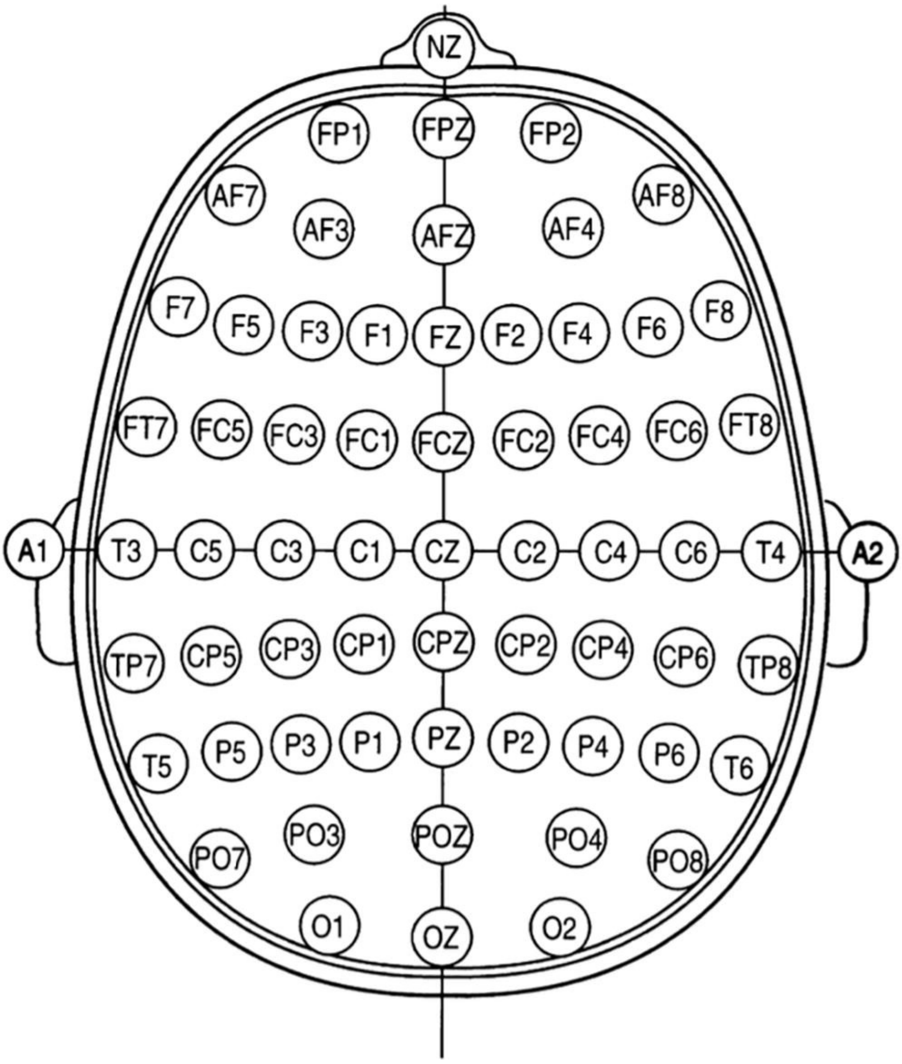

Figure 1.1: Electrode locations of the international 10–10 system for EEG recording (adapted from [1]).

Electroencephalography (EEG) is a prominent non-invasive technique used to record brain activity by placing electrodes on the scalp. A major advantage of EEG is its affordability and portability, making it a practical and accessible tool for both clinical diagnostics and neuroscience research compared to more expensive alternatives like Magnetic Resonance Imaging (MRI). To ensure consistent and reliable data collection, EEG employs the standardized 10–10 system for precise electrode placement. This layout is illustrated in Figure 1.1. This system allows clinicians and researchers to accurately localize brain activity and compare findings across different studies. The combination of this standardized approach with its cost-effectiveness and ease of use enhances its broad applicability. Ultimately, these practical advantages and the dependable data it produces solidify EEG's importance in improving patient care and advancing our understanding of the brain.

An EEG feature is a quantifiable metric computationally derived from the raw brain signal to capture informative aspects of brain activity. Unlike the raw voltage data, features are calculated numerical values, such as spectral power, that represent meaningful neurophysiological information for distinguishing between different states. This process is crucial for reducing the high dimensionality of the original EEG data into a more compact representation. Such a transformation enhances the manageability and robustness of subsequent analyses, including statistical modeling and machine learning (ML) applications.

The recorded EEG signal is characterized by several frequency bands, each associated with specific aspects of brain function. For instance, delta waves (0.5–4 Hz) are linked to deep sleep; theta waves (4–8 Hz) to drowsiness or meditation; alpha waves (8–13 Hz) to relaxed wakefulness; beta waves (13–30 Hz) to active thinking or concentration; and gamma waves (above 30 Hz) to higher cognitive functions and complex processing. Researchers analyze these frequency components to understand transitions between brain states and to examine how neurological conditions alter normal activity.

EEG thus provides valuable data for studying cognitive functions, emotional processing, and the effects of diverse stimuli on brain activity. Its high temporal resolution and capability to capture complex, dynamic electrical patterns have established EEG as a cornerstone of both neuroscience research and clinical neurodiagnostics.

1.1 Focus on the "Richness" principle

Decoding meaning and intent from noninvasive brain signals stands at the crossroads of cognitive neuroscience and brain computer interfaces (BCI). The ambition is to move beyond detecting generic physiological states and toward reading out what a participant perceives or intends at the level of concepts and goals. The friction arises from the way we usually summarize neural data. Conventional pipelines compress rich spatiotemporal activity into low-capacity descriptors such as band-power aggregates, simple temporal statistics, or coarse connectivity indices. These summaries are efficient and useful for quality control, yet they tend to collapse the very relations—across channels, across time, across conditions—that carry semantic regularities. When the scientific target is meaning or intent, representation is not a neutral pre-processing step; it is the first scientific decision about what aspects of neural organization we wish to expose to a decoder.

1.2 Focus on the "Mismatch"

This thesis addresses a specific mismatch between the structure of cognitive intent in the brain and the representational form used in many decoding pipelines. Information regarding intent is relational and context-dependent: it emerges from coordinated patterns distributed over sensors and time and modulated by task demands and prior knowledge. Low-capacity features, typically scalar or short vector outputs and only rarely small matrices, discard much of that relational structure. The result is familiar: modest in-distribution performance, brittle generalization to new subjects or stimuli, and limited interpretability of what, if anything, has been captured about the user's target concept. What is lacking in the literature is a controlled, multi-dataset comparison that treats representational form itself—ranging from low-capacity summaries to structure-preserving matrix- and tensor-like EEG representations—as a primary experimental factor, and that examines how this factor interacts with model class under both in-distribution and transfer-oriented evaluation settings.

1.3 Objectives

The objective of this thesis is to evaluate, in a disciplined and realistic setting, whether EEG representations that retain more spatiotemporal and relational structure support more reliable decoding of task-level meaning and intent. In this thesis, the *Representation Richness Principle (RRP)* states that, for a fixed dataset, task, and computational budget, feature representations that preserve more of the task-relevant spatiotemporal and spectral structure of the EEG signal (e.g., raw waveforms, time–frequency maps, connectivity matrices) tend to support more reliable and more generalizable decoding than heavily compressed scalar or short-vector summaries, provided that the decoding model has sufficient capacity and appropriate inductive biases to exploit this preserved structure.

We refer to the degree to which a representation preserves such multi-dimensional structure as its *representation richness*. Conceptually, we organize the candidate representations along a *richness ladder*: from aggressively compressed scalar or short-vector features , through band-power style vectors, to matrix- and tensor-valued descriptions such as Phase Locking Value (PLV) connectivity matrices, time–frequency maps, and raw multichannel waveforms.

Concretely, we study four EEG representation regimes Raw EEG, Power Spectral Density (PSD), Spectrogram, PLV, and Approximate Entropy (ApEn) across three public datasets and four models (two linear baselines and two deep architectures), using a compact but expressive set of performance metrics.

1.4 Research Questions and Hypotheses

To operationalize the central objective of this thesis—evaluating the impact of representational form on intent decoding—this investigation is guided by three fundamental research questions. These questions are designed to systematically probe the relationship between representational richness and decoding efficacy, considering key factors that determine practical utility. Specifically, they address not only the primary effect on performance but also the critical interactions with model architecture and the crucial test of generalization across different contexts. From these guiding questions, we derive the following specific, testable hypotheses:

- Do richer outputs improve decoding performance compared with low-capacity summaries when evaluated across heterogeneous datasets?
- How does model class moderate these gains—do deep architectures, by virtue of their spatiotemporal and nonlinear inductive biases, translate representational richness into larger improvements than classical baselines?
- Are any observed advantages preserved under transfer—across subjects, sessions, or stimulus sets—where practical BCI typically falter?

Our working hypothesis is that, on average, representations that preserve more of the spatial, spectral, or temporal structure of the EEG signal (Raw, PSD, Spectrogram, PLV) will outperform strongly compressed complexity-based summaries (such as ApEn-only features). We further hypothesize that deep architectures, by virtue of their inductive biases for hierarchical spatiotemporal patterns, will derive larger gains from rich and raw inputs than linear baselines, whereas the latter will remain competitive primarily on lower-dimensional representations.

1.5 Framework

This thesis adopts a pragmatic representational stance, treating each feature map as an explicit hypothesis about how task-relevant information is organized in the brain. Simpler representations, such as band-power vectors or scalar complexity indices, implicitly assume that the target construct can be captured by aggregates or independent channel-wise summaries. Structure-preserving representations, such as time–frequency maps, connectivity matrices, and raw spatiotemporal tensors, instead assume that information resides in relational patterns—correlations, synchrony, and temporal dynamics—that should be retained rather than collapsed. Explicitly symbolic embeddings that align EEG directly with linguistic or conceptual spaces are acknowledged as an important long-term direction, but in this work they serve as conceptual motivation rather than an implemented component of the empirical study. Following this stance, the study contrasts two analysis pipelines. The first reflects a traditional approach in which the signal is pre-processed and reduced to low-capacity features before classification. The second is a richness-aware pipeline that applies models to structure-preserving representations so that their geometric and temporal properties can be exploited directly. Within this framework, model class functions as a moderator of how effectively representational richness can be utilized. Deep learning (DL) architectures with inductive biases for spatiotemporal structure are expected to benefit disproportionately from rich inputs, whereas classical linear models provide conservative, interpretable baselines that are well suited to lower-dimensional features.

1.6 Scope of the Study

This study considers three publicly available EEG datasets with clearly defined task-level objectives: BNCI 2014-001 from Motor Imagery (MI) , BNCI 2014-009 from Event-Related Potential (ERP) and Nakanishi2015 from Steady-State Visually Evoked Potential (SSVEP).

For each dataset, we instantiate four concrete representation regimes: (PSD, Spectrogram, PLV, ApEn) and Raw EEG.

Each representation is evaluated with two models: two linear baselines models are Linear Discriminant Analysis (LDA) and Logistic Regression (LR) and two DL backbones that used Raw EEG signals for input are Convolutional Neural Network (CNN) and Long Short-Term Memory Network (LSTM).

Performance is quantified using a fixed core metric set under subject-aware partitions that support both in-distribution and transfer-oriented evaluation. Subsequent chapters specify architectural details, pre-processing pipelines, and split protocols; this section delineates the empirical scope within which the RRP is examined.

1.7 Dissertation Chapter Overview

This dissertation is organized as follows.

- Chapter 2 (Background) EEG signals encode task-relevant information in a distributed, structured, and time–frequency–dependent way, so representations must preserve this structure rather than rely on a few handcrafted features. The text argues that many strong results with classical or deep models don't generalize, and motivates a "representation-richness" principle: systematically testing which EEG representations actually enable robust, task-independent, and transferable decoding.

- Chapter 3 (Research Design: Datasets, Representations, and Models) sketches the EEG decoding design: three paradigms (ERP, MI, SSVEP) with benchmark datasets, five representations (PSD, PLV, spectrogram, ApEn) and Raw EEG signal, and four models: (LDA, LR, CNN, LSTM) to isolate representation–model effects.

- Chapter 4 (Methodological Framework and Evaluation Protocol) details the complete experimental pipeline, including preprocessing steps, implementation of the five representation regimes, model architectures, training schedules, and the subject-aware validation protocol used throughout.

- Chapter 5 (Experimental Results) reports the empirical results for all representation–model combinations across the three datasets, analyzes representation hierarchies and model-class effects using the fixed core metric set, and summarizes robustness and computational trade-offs.

- Chapter 6 (Discussion) interprets the findings in light of the RRP, neural plausibility, and practical deployment constraints, and revisits the research questions.

- Chapter 7 (Conclusion) consolidates the main contributions, states the final answers to the research questions, and outlines directions for extending representation-rich EEG decoding, including towards explicitly semantic embeddings and multi-view fusion.

Chapter 2

Theoretical Background and Related Work

2.1 Neural Representation of Meaning and Intention

A central premise of this thesis is that meaning and intention are not localized in single neurons or isolated channels, but emerge from distributed, structured patterns of neural activity. Converging evidence from cognitive neuroscience demonstrates that semantic information is encoded within high-dimensional activation spaces.

Mitchell et al. showed that Functional Magnetic Resonance Imaging (fMRI) activity associated with concrete nouns can be predicted using corpus-derived semantic features, indicating that semantically similar words evoke similar spatial activation patterns [2]. Huth et al. further mapped continuous semantic spaces across large regions of cortex during natural speech, revealing smooth topographic organization of meaning rather than isolated category-specific loci [3].

These studies support three key points: (i) semantic content is distributed, (ii) relations between concepts are reflected in geometric relations between neural patterns, and (iii) context and dynamics play a crucial role in representation. When transitioning from fMRI to EEG and BCI, these observations imply that feature extraction procedures for decoding meaning or intention should preserve as much of the relevant spatio–temporal–spectral structure as possible, rather than collapsing the signal into a small set of handcrafted scalars.

In this thesis, we formalize this idea as the RRP : the semantic capacity of an EEG-based representation is directly tied to the degree to which it preserves the structured relationships underlying neural activity.

2.2 Classical EEG Features: Strengths and Structural Limitations

Traditional EEG-based BCI systems are typically built on handcrafted, low-dimensional features, such as:

- band power and PSD in predefined frequency bands;
- time-domain statistics (e.g., mean, variance, Hjorth parameters);
- entropy-based measures (e.g., ApEn, sample entropy);
- spatial filters including Common Spatial Patterns (CSP) for MI.

These approaches have achieved considerable success in controlled applications (e.g., binary MI classification, ERP spellers, SSVEP-based selection, seizure detection), largely due to their:

- computational efficiency and suitability for real-time systems;
- interpretability and physiological plausibility;
- compatibility with simple linear classifiers such as LDA and LR.

However, they are intrinsically reductive. Temporal averaging assumes local stationarity and removes fine temporal dynamics; spatial projections reduce the multichannel structure to a few components; marginalization over broad frequency bands discards cross-frequency patterns and harmonic structure. From the perspective of distributed semantic coding, such operations collapse the underlying representational geometry into a small set of scalars, potentially eliminating precisely those relationships that encode latent cognitive and semantic content.

Large-scale benchmarking frameworks such as MOABB (Mother of All BCI Benchmarks) have systematically demonstrated that many pipelines based on conventional features exhibit strong dataset dependence and poor cross-dataset generalization, even within the same paradigm [4]. Methods that appear state-of-the-art on a single dataset often fail to reproduce their performance under different recording conditions, subject populations, or experimental designs. This motivates a more principled analysis of how the choice of representation constrains generalization.

2.3 Structure-Preserving Representations of EEG

To address the limitations of heavily compressed features, several families of representations seek to preserve more of the intrinsic structure of EEG data.

2.3.1 Time–Frequency Representations

Time–frequency methods, such as the short-time Fourier transform and wavelet transforms, produce two-dimensional maps of power or amplitude over time and frequency. These spectro-temporal representations capture transient oscillatory events and non-stationary dynamics that are invisible to stationary PSD estimates, and they are naturally compatible with convolutional architectures that exploit local structure.

2.3.2 Covariance Matrices and Riemannian Geometry

Spatial covariance matrices summarize pairwise relationships between EEG channels and have proven highly effective in MI classification and related tasks. Treating these matrices as points on the Riemannian manifold of symmetric positive-definite (SPD) matrices yields features and classifiers with improved robustness compared to naïve Euclidean treatments [4]. This line of work explicitly acknowledges that EEG is a structured object, and leverages this structure rather than discarding it.

2.3.3 Connectivity and Phase-Based Measures

Measures such as PLV, coherence, and related connectivity metrics encode interaction patterns between brain regions. Representations based on connectivity matrices or graphs capture network-level organization, which is hypothesized to be central to cognitive and semantic processing. These representations therefore provide another route toward structure-preserving characterization of EEG activity.

2.3.4 High-Order and Raw-Signal Representations

Higher-order tensorial representations and the direct use of multi-channel raw EEG preserve maximal information. In principle, such inputs allow learning algorithms to discover relevant patterns including cross-channel dependencies, temporal motifs, and non-linear relationships without being restricted by strong, manually imposed compression.

2.4 Model Classes and the Representation–Model Interaction

The effectiveness of a representation cannot be separated from the choice of model. A representation that is too simple constrains any model, while a rich representation paired with an inadequate model may fail to realize its potential.

2.4.1 ML Models

LDA, LR, and support vector machines (SVMs) have long been standard in BCI. They are well-suited to low-dimensional, engineered features, relatively robust in small-sample regimes, and straightforward to interpret and deploy. Nevertheless, their reliance on comparatively simple decision boundaries may be misaligned with the complex manifolds that EEG data occupy.

2.4.2 DL Models

DL introduces architectures that exploit the structure of EEG representations directly. Schirrmeister et al. demonstrated that CNNs applied end-to-end to raw EEG can achieve competitive or superior performance compared to traditional pipelines, while also offering insight into learned features [5]. Lawhern et al. proposed EEGNet, a compact convolutional architecture designed for multiple paradigms (ERP, Error-Related Negativity (ERN), Movement-Related Cortical Potential (MRCP), and Sensorimotor Rhythm (SMR)) with good performance and low parameter count [6].

These models leverage local receptive fields, hierarchical feature abstraction, and non-linear transformations that naturally align with structured inputs (e.g., raw EEG, spectrograms, topographic maps). However, reproducibility and cross-dataset studies have shown that deep models can overfit to specific datasets, pre-processing choices, and hyperparameters; when evaluated under rigorous subject-wise and dataset-wise splits, their generalization is often more modest than initial within-dataset results suggest [7].

These observations reinforce the central hypothesis of this thesis: performance and generalization emerge from the interaction between representation richness and model capacity. Neither can be evaluated in isolation.

2.5 Empirical Evidence of Limited Generalization in BCI

Recent studies provide concrete examples where highly sophisticated pipelines achieve outstanding performance on specific datasets, yet fail to constitute universal solutions. These examples are directly relevant for the motivation of the present work.

2.5.1 High-Accuracy Ensembles for MI

Ahmadi and Mesin introduced a correlation-optimized weighted stacking ensemble (COWSE) for MI EEG classification, combining multiple base classifiers with correlation-informed stacking to achieve very high accuracies on standard MI datasets [8]. In a follow-up IEEE Access paper, they proposed a weighted and stacked adaptive integrated ensemble model evaluated across multiple MI datasets, again reporting strong performance [9].

These studies highlight the power of carefully designed ensembles tuned to specific data. However, when the same architectures and feature configurations are directly applied to other MI datasets—even within the same paradigm—the reported performance levels are not consistently reproduced. This discrepancy is consistent with MOABB findings and underscores that such pipelines, while excellent locally, do not yet resolve the broader challenge of robust, cross-dataset MI decoding.

2.5.2 Hierarchical Deep Models for Coma Outcome Prediction

In another study a hierarchical binary classification framework that proposed for CNN-based feature extraction with traditional methods coma outcome prediction using EEG [10]. The reported performance on the target dataset is near perfect.

While clinically promising, attempts to reuse the same hierarchical architecture on different cohorts or recording setups yield more variable results. As in MI, this suggests that extremely high within-dataset accuracy can reflect tight coupling to dataset-specific characteristics rather than universally stable EEG markers.

2.5.3 Adversarial Training for Secure Brain-to-Brain Communication

The same pattern appears in the emerging area of secure BCIs and brain-to-brain communication. Ahmadi, Kuhestani, and Mesin proposed an adversarial training framework to secure brain-to-brain communication channels, demonstrating strong robustness to adversarial perturbations on their datasets [11]. Subsequently, adversarial training was applied to SSVEP-based EEG signals to secure communication channels with similarly promising results [12].

However, direct transfer of these architectures and training schemes to different SSVEP datasets or alternative recording protocols does not always maintain the same security and performance guarantees. Again, the effectiveness of the pipeline is contingent on specific data properties.

2.5.4 Synthesis: Lack of a General Design Principle

Across MI ensembles, coma outcome prediction, and secure BCI frameworks, a coherent picture emerges:

- reported performances are often exceptional on the original datasets.
- re-using the same pipelines on different datasets or paradigms yields inconsistent results.
- there is no established, principled mapping from EEG data type or task to an "optimal" set of features and models.

These observations, supported by large-scale benchmarks and reproducibility studies, indicate that current BCI methodology largely operates in a dataset-specific regime and lacks a generalizable theory of representation and modeling. Addressing this gap is one of the central aims of the present thesis.

2.6 Towards Task-Independent and Semantic EEG Representations

Recent work has explicitly targeted the development of EEG representations that are both task-independent and semantically meaningful, moving beyond narrow, protocol-specific feature engineering.

2.6.1 Topographic Sequence Representations

In "Decoding Visual Imagination and Perception from EEG via Topomap Sequences", Ahmadi and Mesin propose transforming each EEG trial into a sequence of scalp topographic maps, which is then processed as a spatio–temporal image sequence [13]. This approach:

- preserves spatial distributions over electrodes at each time point;
- models the temporal evolution of activity explicitly;
- is conceptually compatible with multiple visual and cognitive paradigms.

By avoiding aggressive scalar compression, this framework is more consistent with the notion of distributed coding and constitutes a step toward more general-purpose representations.

2.6.2 Universal Semantic Feature Extraction

The work “Universal Semantic Feature Extraction from EEG Signals: A Task-Independent Framework” introduces, for the first time, a systematic approach for learning a universal semantic embedding of EEG signals [14]. The proposed framework integrates deep architectures to:

- capture low-level spatio–temporal patterns;
- learn higher-level, task-agnostic semantic features shared across datasets;
- outperform conventional feature sets on multiple evaluation tasks.

Conceptually, this framework operationalizes the hypothesis that a common representational space can organize diverse EEG tasks (e.g., MI, perception, higher cognitive functions), bringing the field closer to task-independent decoding.

2.6.3 Implications

Topomap sequences and universal semantic feature extraction exemplify a new generation of representations that:

- respect the distributed and structured nature of neural activity.
- reduce reliance on ad-hoc, task-specific features.
- offer promising candidates for higher semantic capacity and improved cross-dataset robustness.

Nonetheless, these approaches have primarily been evaluated on limited sets of datasets and tasks. Their advantages over classical and intermediate representations must be assessed under rigorous, standardized protocols to determine where and how they truly generalize.

2.7 Summary and Positioning of the Present Work

The literature reviewed in this chapter leads to the following conclusions:

- Neural meaning is distributed and structured. Semantic information is encoded in high-dimensional, relational patterns, implying that EEG features for decoding meaning or intention should preserve structural relationships rather than collapse them.
- Classical EEG features are useful but inherently limited. They support simple and interpretable BCIs under constrained conditions but discard much of the spatio–temporal organization likely required for robust semantic decoding and exhibit limited cross-dataset generalization.

- Structure-preserving representations and deep models provide richer alternatives, but their robustness is not guaranteed. Improved performance on individual datasets does not automatically translate into broad generalization.
- Recent task-independent and semantic frameworks (e.g., topomap sequences, universal semantic features) are promising but under-examined. Their relationship to simpler baselines under strict evaluation settings remains to be systematically clarified.
- There is currently no generally accepted design principle connecting data type, representation, and model choice in EEG-based decoding.

Motivated by these points, the present thesis formulates and empirically tests the RRP for EEG. Specifically, we:

- define a hierarchy of representations, from traditional PSD and entropy-based features to spectrograms, PLV/connectivity, and raw EEG;
- combine these representations with both classical (e.g., LDA, LR) and deep architectures (e.g., CNN-based models);
- evaluate their performance and generalization under rigorous cross-subject and cross-dataset protocols.

By doing so, this work aims to move beyond isolated best-case reports and toward a principled understanding of which kinds of EEG representations truly capture stable, transferable aspects of underlying neural computations and semantic content.

Chapter 3

Research Design: Datasets, Representations, and Models

3.1 Introduction: Conceptual Architecture of the Empirical Investigation

The feature set is chosen to probe complementary hypotheses about where task-relevant structure is expressed in EEG signals. Each representation offers a distinct inductive view on the same underlying data, enabling a controlled comparison without committing to a single, narrow prior, the signal representations that encode competing theoretical hypotheses regarding the locus of meaningful information in EEG signals, and the model families that afford varying capacities for learning and generalization. The objective herein is to articulate the *what* and *why* of our design decisions, reserving the technical elaboration of *how* these elements are implemented for chapter 4.

3.2 Paradigms and Target Signals

The empirical scope of this thesis centers on three well-established EEG paradigms that provide distinct target constructs for decoding. Each paradigm is defined by its elicitation procedure and canonical neural signature, which together determine the structure of the data and the nature of the labels analyzed in subsequent chapters. A concise synopsis appears in Table 3.1; detailed acquisition protocols and dataset-specific settings are provided in Chapter 4, in particular Sections 4.2 and 4.3.

- **ERP paradigm.** The ERP paradigm centers on a positive deflection typically observed approximately 250–500 ms after infrequent, task-relevant stimuli in oddball paradigms. It is commonly interpreted as reflecting context updating, allocation of attentional resources, and stimulus evaluation. In canonical BCI applications, the decoding target is whether a presented item is the attended target (target vs. non-target). Practical challenges include class imbalance inherent to oddball designs, temporal jitter in the evoked response, and pronounced inter-individual variability in both latency and amplitude. Paradigm-specific acquisition schemes and preprocessing steps for all ERP recordings used in this thesis are detailed in Sections 4.2 and 4.3 of Chapter 4.

- **MI paradigm.** involves covert rehearsal of limb or tongue movements without overt execution. The paradigm is associated with systematic modulations of oscillatory activity over sensorimotor areas—classically described as event-related desynchronization/synchronization in the μ/β ranges—and typically yields multi-class labels reflecting the imagined effector (e.g., left/right hand, feet, tongue). Common challenges include strong inter-subject differences in spatial activation patterns, non-stationarity across sessions, and variability in imagery strategy, vividness, and compliance. The recording configurations, cueing schemes, and preprocessing pipeline for the MI paradigm are specified in Sections 4.2 and 4.3 of Chapter 4.

- **SSVEP paradigm.** present periodic visual stimulation at distinct frequencies; the ensuing neural response in visual cortex is phase-locked to the driving frequency and its harmonics. The decoding target is the frequency corresponding to the attended stimulus among a discrete set of alternatives. Typical challenges include close spacing between candidate frequencies, susceptibility to visual fatigue and reduced vigilance over time, and subject-specific differences in resonance properties and harmonic content. The SSVEP stimulation layouts, frequency allocations, and associated preprocessing steps are summarized in Sections 4.2 and 4.3 of Chapter 4.

Paradigm	Target Construct	Neural Signature	Challenges
ERP	Target vs. non-target decision in oddball tasks	Positive ERP ∼250–500 ms post-stimulus; time-locked deflection reflecting context updating	Class imbalance; latency jitter; inter-subject variability
MI	Imagined effector (e.g., L/R hand, feet, tongue)	Oscillatory modulation over sensorimotor cortex; ERD/ERS in μ/β bands	Inter-subject spatial variability; session non-stationarity; variable imagery strategies
SSVEP	Attended stimulus among discrete flicker frequencies	Phase-locked steady-state response at driving frequency and harmonics in occipital areas	Close-spaced frequencies; visual fatigue; subject-specific resonance profiles

Table 3.1: Summary of EEG paradigms, their target constructs, canonical neural signatures, and typical challenges.

3.3 Datasets Preview

The choice of datasets is critical for ensuring the validity, reproducibility, and interpretability of the empirical findings. This study employs three well-established public datasets that are widely used as benchmarks for BCI research. Each dataset instantiates one of the paradigms introduced in Section 3.2, thereby covering distinct forms of task-relevant or "semantic" information to be decoded. A compact visual summary of their core characteristics is provided in Figure 3.1, which is referenced throughout this chapter when discussing paradigm-specific design choices and evaluation constraints.

- **BNCI 2014-001 (MI).** This dataset is a canonical benchmark for decoding motor-related intentions from non-invasive EEG. It comprises recordings from 9 healthy subjects (aged approximately 20–30 years), using 22 EEG and 3 Electrooculography (EOG) channels at a sampling rate of 250 Hz. Two sessions were recorded on different days for each subject; each session consists of 6 runs, and each run contains 48 trials (12 per class), yielding 288 trials per session. The four-class setup distinguishes imagined movements of the left hand, right hand, both feet, and tongue, providing a structured yet challenging multi-class problem with realistic inter-subject variability.

- **BNCI 2014-009 (ERP).** This dataset serves as a representative benchmark for ERP decoding in a visual ERP speller paradigm, operationalizing the detection of attended versus non-attended matrix elements. It contains data from multiple healthy subjects performing row/column-based spelling with repeated brief flashes of stimulus groups. Across overt and covert attention conditions, several sessions per subject are recorded, including structured spelling runs and free-spelling runs. The resulting data exhibit strong class imbalance (few target trials vs. many non-target trials) and trial-to-trial variability in the elicited potentials, making it well suited to stress-test robustness to low signal-to-noise ratio (SNR) and unequal class distributions.

- **Nakanishi2015 (SSVEP).** This dataset is a cornerstone benchmark for SSVEP decoding in high-speed speller interfaces. It includes recordings from 10 healthy subjects who attend to one of 40 simultaneously presented visual stimuli, each tagged by a distinct frequency–phase combination in the approximate 8–15.8 Hz range. EEG is recorded from occipital and parietal areas using a high-density montage over multiple runs. The large target set and fine spacing between stimulation frequencies impose a stringent requirement on frequency-specific and subject-robust decoding, aligning closely with real-world demands for high information transfer rates.

By jointly considering these three complementary benchmarks, this thesis evaluates decoding performance across transient, oscillatory, and steady-state response regimes. This design supports more generalizable conclusions about how different representational choices interact with the underlying neural dynamics and task structure, rather than overfitting insights to a single paradigm.

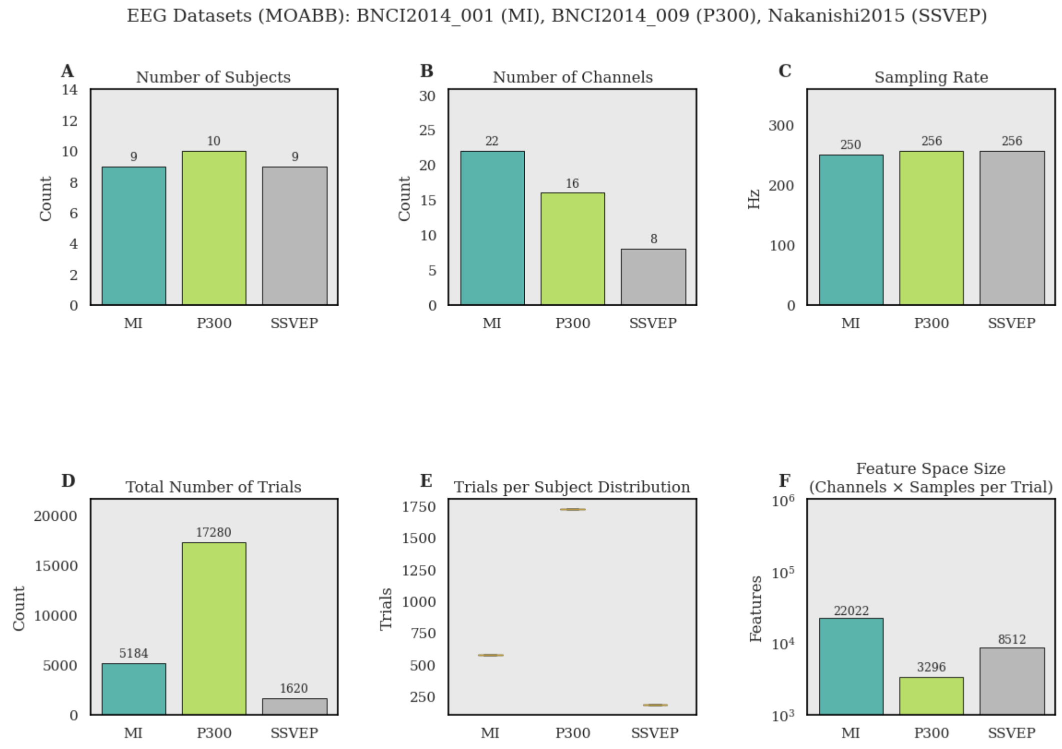

Figure 3.1: Overview of the three benchmark datasets used in this study, including number of subjects, channels, sampling rate, and representative normalized EEG segments for each paradigm.(A)Total number of subjects in each paradigm across all subjects available in MOABB. The value above each bar shows the exact count.(B)Number of EEG channels actually used in the analysis of each dataset after paradigm-specific epoching.(C)Effective sampling rate of epochs after MOABB preprocessing.(D)Total number of trials/epochs aggregated over all subjects for each paradigm.(E)Distribution of the number of trials per subject.The center line indicates the median.(F)Size of the raw feature space after epoching; the product of the number of channels and samples per epoch. A logarithmic axis is used to make order-of-magnitude differences clearer.

3.3.1 Overview of Feature Representations

This thesis focuses on four complementary feature representations that capture distinct aspects of EEG dynamics: spectral power, time–frequency structure, phase relationships, and signal complexity. Together, they provide a structured way of probing how different facets of neural activity contribute to robust decoding.

- PSD quantifies how the signal's power is distributed across frequencies. By decomposing the EEG into its constituent frequency components, PSD highlights rhythmic activity such as α, β, or γ bands and allows the extraction of band-limited power profiles. As a stationary frequency-domain descriptor, PSD is straightforward to compute and interpret, and serves as a natural baseline for characterizing oscillatory structure in neural signals.

- PLV measures the consistency of phase differences between two signals across trials or time. Instead of focusing on power at individual electrodes, PLV captures the stability of their phase relationships, providing a proxy for functional connectivity or synchronization between brain regions. High PLV indicates tightly coordinated activity, whereas low PLV reflects more independent or desynchronized signals.

- Spectrogram-based representations describe how the spectral content of the signal evolves over time. By computing frequency-specific power within short, sliding windows (e.g., via short-time Fourier or wavelet transforms), they provide a time–frequency map that preserves both temporal and spectral structure. This representation is particularly useful when relevant information is carried by transient or non-stationary patterns that cannot be captured by a single, global spectrum.

- ApEn is a non-linear complexity measure that quantifies the regularity and predictability of a time series. Lower ApEn values indicate more regular and repetitive dynamics, whereas higher values reflect greater irregularity. As a compact descriptor of temporal structure, ApEn is sensitive to subtle changes in signal organization and can complement classical linear features by emphasizing the degree of underlying neural complexity.

Taken together, these four representations span power-based, connectivity-based, time frequency, and complexity-oriented perspectives on EEG activity, enabling a systematic comparison of how different facets of the signal support reliable decoding.

3.4 Model Families

The primary objective of this study is to evaluate the representational richness of EEG feature representations, rather than to identify a single "best" classifier. Models are therefore treated as controlled analysis tools: each model family provides a distinct inductive bias, and performance differences are interpreted in relation to how effectively a given representation exposes task-relevant structure. In this way, conclusions remain centered on representations and their interaction with model families, not on architecture-specific fine-tuning.

To preserve comparability, a small set of widely used and conceptually transparent models is selected. All models are applied under matched evaluation protocols, subject-aware data splits, and fixed hyperparameters across datasets and feature types (see Section 4.4 in Chapter 4).

This constrained design reduces the risk of overfitting model choices to individual datasets and allows performance patterns to be read as evidence about representational suitability. Table 3.2 summarizes the four models considered in this thesis.

Two are classical ML baselines operating on vectorized feature representations; two are DL architectures designed to exploit the spatial and temporal structure of EEG data. Together, they span low-capacity linear decision rules and flexible non-linear function approximators, providing a coherent framework for analyzing how different feature spaces interact with different inductive biases.

3.4.1 ML Baselines

In this thesis, *ML* refers to classical discriminative models trained on pre-computed feature vectors. These models are computationally efficient, data-economical, and offer interpretable decision boundaries, making them natural reference points for assessing whether more expressive representations or architectures are actually needed.

- **LDA Model.** learns a linear projection that maximizes the ratio of between-class to within-class variance under Gaussian class assumptions with shared covariance. It is widely used in BCI research due to its simplicity, robustness in low-data regimes, and fast inference. In this study, LDA serves as a low-capacity baseline that tests whether a given representation already organizes the data into approximately linearly separable clusters.
- **LR Model.** class membership via linear decision functions passed through a softmax, yielding well-calibrated probabilistic outputs. Applying ℓ_2-regularization controls overfitting and stabilizes the weight estimates, especially in high-dimensional feature spaces. Here, regularized LR provides an interpretable probabilistic baseline that can exploit subtle but linearly decodable structure in the feature representations.

3.4.2 DL Backbones

In this thesis, DL denotes neural network architectures that operate directly on structured EEG inputs (e.g., sequences, time–frequency maps) and learn hierarchical non-linear feature transformations. These models introduce stronger inductive biases for spatial and temporal patterns and can, in principle, discover task-relevant structure beyond what is explicitly encoded in hand-crafted features.

- **CNN Model.** applies temporal and/or spatio–temporal convolutions to EEG inputs to learn localized filters that detect characteristic patterns across channels and time. Stacking convolutional and pooling layers yields progressively more abstract representations while keeping the number of parameters controlled. In this work, the CNN is used as a generic end-to-end architecture to test whether structured inputs allow non-linear filters to extract additional discriminative information beyond linear baselines.

- **LSTM Model.** EEG as a sequence and uses recurrent units with gating mechanisms to capture temporal dependencies over multiple time scales. This allows the model to integrate information across extended windows and to represent temporal context that may be relevant for decoding. Here, the LSTM serves as a complementary deep model with an explicit sequence-processing bias, enabling a comparison between convolutional and recurrent approaches under the same evaluation protocol.

Category	**Model**	**Core idea**	**Data/Compu**	**Interpret.**	**Typical BCI use**
ML	LDA	Linear projection maximizing separation between class means under shared covariance	Low	High	Canonical baseline for ERP and motor imagery decoding
ML	LR	Linear decision boundaries with probabilistic outputs and ℓ_2-regularization	Low	High	Robust baseline for binary and multi-class BCI tasks
DL	CNN	Hierarchical feature extraction via convolutions on structured EEG inputs	Medium–High	Moderate	End-to-end decoding across diverse EEG paradigms
DL	LSTM	Recurrent modeling of temporal dependencies in EEG sequences	Medium–High	Low–Moderate	Tasks with pronounced temporal structure and sequential dynamics

Table 3.2: Summary of the four model instances used in this thesis.

3.5 Rationale for the Selection of Feature Representations

The feature set is chosen to probe complementary hypotheses about where task-relevant structure is expressed in EEG signals. Each representation offers a distinct inductive view on the same underlying data, enabling a controlled comparison without committing to a single, narrow prior.

- PSD encodes the distribution of signal power across frequencies and tests whether stable band-specific activity is sufficient to separate conditions. It provides a compact, well-understood representation that prioritizes spectral composition over precise temporal evolution or inter-channel relations.

- Spectrogram preserving how spectral content evolves over time. They operationalize the hypothesis that relevant information is carried by transient or non-stationary patterns, without reducing the signal to a single global spectrum.

- PLV captures the stability of phase differences between channels and thus targets information expressed in coordinated activity rather than in individual amplitudes. It embodies a connectivity-centric view in which task-relevant structure is reflected by consistent coupling within functional networks.

- ApEn summarizes the regularity of the signal and tests whether changes in cognitive or sensorimotor state manifest as shifts in complexity or predictability. It complements power- and connectivity-based views with a compact, non-linear descriptor of temporal organization.

- Raw inputs retain the original spatiotemporal resolution with minimal preprocessing and impose the weakest hand-crafted prior. They instantiate the hypothesis that suitable architectures can recover task-relevant structure directly, providing a reference against which the added value of engineered features can be assessed.

Taken together, these representations span spectral, time–frequency, connectivity, complexity, and end-to-end views of the data, allowing subsequent analyses to attribute observed performance patterns to specific assumptions about how EEG encodes task information.

3.6 Rationale for the Selection of Models

We selected two representative models from both the ML and DL families. However, we additionally carried out experiments on several other similar models and observed consistent results.

The model set is deliberately minimal and structured to support representation-centered conclusions. Rather than optimizing over many architectures, this study employs a small number of stable, widely used models with complementary inductive biases and comparable capacity constraints. Their role is to act as probing instruments for the feature spaces defined above.

Two models instantiate classical *ML* on fixed feature vectors; two models instantiate *DL* on structured inputs. This pairing provides a controlled contrast between shallow linear decision boundaries and flexible non-linear function approximators. An overview of their roles and properties is provided in Table 3.2.

- **LDA and LR.** These linear models represent the conventional BCI pipeline of hand-crafted features followed by low-variance decoders. They are computationally efficient and interpretable, and their performance directly reflects how linearly accessible task-relevant information is within each engineered feature space.
- **CNN.** The CNN backbone introduces a structured, locality-aware non-linearity that can exploit spatial and temporal patterns in raw or transformed inputs. Under a capacity-controlled configuration, it probes whether richer input structure allows additional discriminative information to be extracted beyond what linear baselines can capture.
- **LSTM.** The LSTM backbone focuses on sequential dependencies by integrating information over time via gated recurrent units. It serves as a complementary deep architecture that tests whether explicitly modeling temporal context yields further gains from the same inputs.

Across all experiments, these four models are trained under shared, conservative protocols. This design ensures that differences in performance can be interpreted primarily as interactions between feature representations and model families, rather than as artifacts of aggressive architecture tuning.

Collectively, this chapter has specified the conceptual architecture of the empirical study. Three paradigms (ERP, MI, SSVEP) were introduced as distinct targets for decoding; three benchmark datasets were selected to instantiate these paradigms in a controlled and widely reproducible manner; a compact set of feature representations was defined to encode complementary hypotheses about how task-relevant information is expressed in EEG signals; and four model instances were chosen to probe these representations under transparent and capacity-controlled inductive biases. This design constrains interpretability: observed performance patterns in later chapters can be systematically attributed to interactions between paradigms, representations, and model families, rather than to ad hoc architectural choices or dataset-specific tuning.

Chapter 4 now operationalizes this blueprint. It details the concrete preprocessing pipelines, epoching schemes, normalization steps, and implementation settings for each paradigm and dataset; formalizes the training, validation, and subject-aware evaluation protocols shared across all configurations; and specifies the metrics and statistical procedures used to assess and compare performance. In doing so, it provides the methodological instantiation of the research design articulated here, ensuring that the subsequent empirical results can be traced back to clearly defined and reproducible analytical decisions.

Chapter 4

Methodological Framework and Evaluation Protocol

Five representation pathways—Raw, PSD, Spectrogram , PLV , and ApEn are compared across three canonical paradigms (MI, ERP, SSVEP) under two modeling scenarios (ML and DL). To support a fair, imbalance-robust comparison aligned to Section 1.2, we lock a core set of four discrimination metrics used consistently across all experiments and both scenarios Figure 4.1.

We explicitly decouple three axes that are often entangled in the EEG literature: (i) representation richness, i.e., what the input encodes (scalar vectors vs. matrices/tensors, band-limited vs. broadband); (ii) model class, i.e., linear discriminants vs. parametric sequence/convolutional learners; and (iii) evaluation stance, i.e., per-subject aggregation under subject-aware splitting with robust metrics and paired tests. By keeping (ii) and (iii) fixed across all (i), we can attribute performance differences primarily to representation rather than to accidental strengths of one pipeline.

4.1 What "richness" means in this study

Building on the notion of representation richness and the richness ladder introduced in Section 1.3, we treat richness as a first-class experimental factor rather than a byproduct of engineering convenience. In this chapter, the categories are instantiated as follows:

- **Low-capacity vectors.** Per-channel or per-band summaries (e.g., PSD bins, ApEn values) that compress temporal structure aggressively. They offer favourable SNR in small-data regimes but risk discarding cross-channel and temporal context that may carry semantic content.
- **Matrix/tensor views.** Representations such as PLV channel–channel matrices or Spectrogram $F \times \tau$ maps (optionally per channel) that preserve relational and temporal geometry at the cost of higher dimension and potentially greater

variance. These views can capture phenomena such as induced synchrony or tagging harmonics (SSVEP) when properly regularized.

- **Raw.** Multichannel waveforms that preserve maximal temporal fidelity with minimal priors, delegating representational learning to the model and providing a neutral baseline against which engineered features can be fairly evaluated.

To keep the causal reading of results clean, we deliberately avoid multi-view fusion at this stage. Each representation is evaluated in isolation to answer: "What can this representation support on its own, under matched budgets and protocols?"

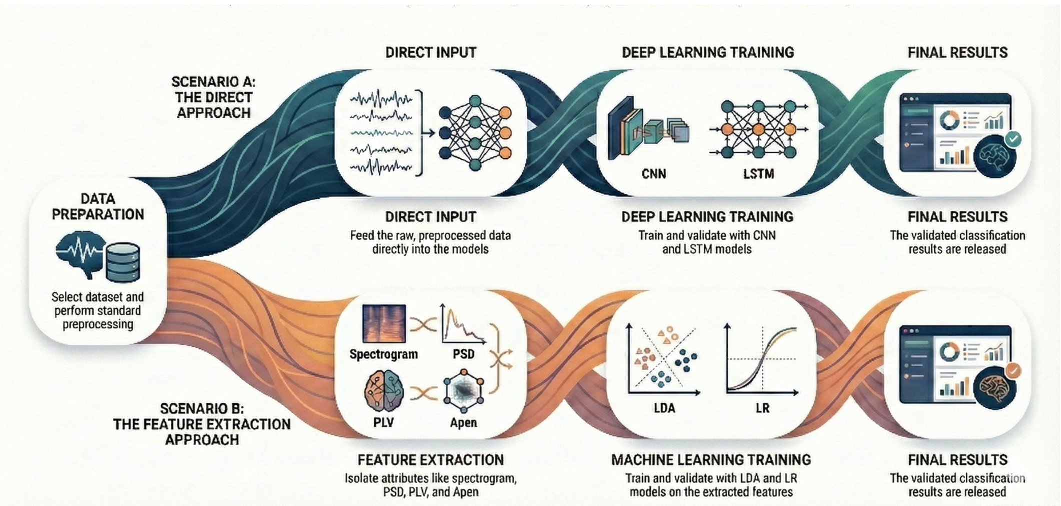
SCENARIO A: THE DIRECT APPROACH
DIRECT INPUT
DEEP LEARNING TRAINING
FINAL RESULTS
CNN
LSTM
DATA PREPARATION
Select dataset and perform standard preprocessing
DIRECT INPUT
Feed the raw, preprocessed data directly into the models
DEEP LEARNING TRAINING
Train and validate with CNN and LSTM models
FINAL RESULTS
The validated classification results are released
Spectrogram
PSD
PLV
Apen
LDA
LR
SCENARIO B: THE FEATURE EXTRACTION APPROACH
FEATURE EXTRACTION
Isolate attributes like spectrogram, PSD, PLV, and Apen
MACHINE LEARNING TRAINING
Train and validate with LDA and LR models on the extracted features
FINAL RESULTS
The validated classification results are released

Figure 4.1: The experimental framework illustrating the dual-path analysis Methodology

4.2 Data Loading with MOABB

4.2.1 Environment and library roles

Reproducibility stands as a cornerstone of scientific inquiry, yet it has historically been a significant challenge in the field of BCI research due to variations in data handling, preprocessing pipelines, and evaluation methodologies. To address this directly and to build our framework on a foundation of transparency and standardization, we leverage the MOABB framework. MOABB, which is built atop MNE-Python, the powerful and widely adopted Python library for neurophysiological data analysis, provides a crucial layer of abstraction. It allows for consistent, programmatic access to a wide range of public BCI datasets and defines standardized evaluation paradigms. This approach is instrumental in our goal of minimizing implementation-induced variance—subtle differences in code that can lead to divergent results—and thereby facilitating fair and meaningful cross-study comparisons. MOABB ensures that when we instantiate a specific paradigm, such as MI, it automatically applies standardized choices for temporal apertures (the time window of interest for each trial) and band-pass frequency ranges that are well-established in the literature as being physiologically relevant to the task. This systematic approach prevents inconsistencies and ensures that our results are comparable not only within this study but also with the broader scientific community.

4.2.2 Deterministic mapping from dataset to paradigm

To prevent silent drift, the paradigm is inferred deterministically from the dataset key as summarized in Table 4.1.

This mapping triggers paradigm-specific time windows and passbands at instantiation time, not via global or hidden defaults. The centralized dictionary `PARADIGM_CONFIGS` is the single source of truth (SST) for temporal apertures ($t_{\min}, t_{\max}$), band-passes ($f_{\min}, f_{\max}$), sampling rates, and class labels.

Paradigm	Dataset	#Classes	Rate (Hz)	Window (s)	Band-pass (Hz)
MI	`BNCI2014_001`	4	250	0.0–4.0	7–35
ERP	`BNCI2014_009`	2	256	0.0–1.0	1–24
SSVEP	`Nakanishi2015`	12	256	0.0–4.0	1–50

Table 4.1: Paradigm configurations provided to MOABB at instantiation time.

4.2.3 Acquisition sequence and output format

Loading proceeds as: (i) dataset instance → (ii) paradigm instance with explicit injection of Table 4.1 → (iii) subject list from the dataset → (iv) `paradigm.get_data(dataset, subjects)`. We return a triple $(\mathbf{X}, \mathbf{y}, \text{metadata})$ where

$$\mathbf{X} \in \mathbb{R}^{N \times C \times T}, \qquad \mathbf{y} \in \{0, \ldots, K-1\}^N, \qquad \text{metadata} \ni \text{subject ids, run/session.}$$

String labels are *LabelEncoded* to contiguous integers and cast to `int64`. Subject IDs are extracted at the epoch level and preserved for subject-aware splitting (§4.3.5). Immediately after loading, we log shapes, label support, and dataset sampling rate; malformed or NaN/Inf trials are flagged for removal in preprocessing.

4.3 Preprocessing: Contracts, Leakage Control, and Numerical Stability

The preprocessing layer upholds five invariants to ensure consistency, prevent data leakage, and enhance numerical stability. It standardizes tensor formats with channels-first (N, C, T) for time-domain and (N, C, F,) for time-frequency data. Normalization and scaling are fitted solely on training windows per fold and fixed for validation/test sets. Raw data receives minimal conditioning via per-channel demeaning, avoiding mandatory filtering for spectral neutrality. Features needing band limits, like PSD or PLV, use high-precision float64 filtering before float32 storage. Partitioning respects subjects and windows, preventing cross-fold straddling to maintain integrity.

4.3.1 Guiding principles and guarantees

Our preprocessing layer guarantees five invariants:

- **Unified tensor contracts:** To facilitate fair model swapping and comparison, inputs are standardized into specific tensor formats: time-domain data, such as raw EEG, uses a 3D structure $\mathbf{X} \in \mathbb{R}^{N\times C\times T}$ representing trials by channels by time points, while time-frequency data like spectrograms employs a 4D format $\mathbf{X} \in \mathbb{R}^{N\times C\times F\times \tau}$ for trials by channels by frequencies by time windows. All data is converted to float32 for efficiency, with tensors arranged in a memory-contiguous, channels-first order to boost performance in DL environments.

- **Train-only estimation.** This is the cornerstone of our anti-leakage strategy. Any transformation that learns parameters from the data, such as a standardization scaler or a dimensionality reduction component, is fitted *only* on the training data windows within a given cross-validation fold. The parameters learned from the training data (e.g., mean and standard deviation) are then applied, without modification, to the validation and test windows of that same fold. This rigorously simulates a real-world scenario where the model has no knowledge of future data.

- **Minimal conditioning for Raw.** The "Raw" representation path is intended to serve as a spectrally neutral baseline against which more complex, engineered features can be judged. To preserve this neutrality, we apply only a per-channel demean (a constant detrend). Crucially, no compulsory band-pass filtering is imposed on the Raw pathway. This is a deliberate choice to avoid pre-supposing which frequency bands are important, allowing the DL models, in particular, to learn relevant spectral features from the wideband signal directly.

- **Feature-targeted filtering.** In contrast to the Raw pathway, when band-limiting is an intrinsic and necessary component for a feature's calculation (e.g., for calculating PSD within specific bands or for estimating band-specific PLV), we utilize well-defined filter banks. Sensitive numerical operations, such as applying Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filters and the Hilbert transform for phase estimation, are performed using high-precision 64-bit floating-point arithmetic float64 to minimize numerical errors before the results are safely downcast to float32.

- **Subject-aware partitioning.** Subject leakage arises when data from the same individual is present in both training and test sets, enabling models to memorize subject-specific idiosyncrasies rather than learn population-level patterns. To avoid this, data must be partitioned strictly at the subject level—assigning each individual's complete recordings to a single split (train, validation, or test) within each fold—thereby preventing cross-partition overlap and promoting genuinely generalizable model performance.

4.3.2 Harmonization and demean

Upon initial loading, the EEG channels are harmonized to a fixed 10-20 system ordering. This ensures that the channel dimension (C) has a consistent anatomical meaning across all datasets and models, which is particularly important for spatially-aware models like CNNs. Any non-EEG channels, such as EOG or ECG, are dropped unless they are explicitly required for a specific artifact removal algorithm (which is not used by default in our minimal preprocessing approach). The continuous data is then segmented into trials using the paradigm-dependent temporal apertures defined in Table 4.1. Immediately after epoching, each trial undergoes a per-channel demeaning operation. This simple yet effective step subtracts the mean value of the time series for each channel within that specific trial:

$$\tilde{x}_{c,t} = x_{c,t} - \frac{1}{T_{\text{trial}}} \sum_{u=1}^{T_{\text{trial}}} x_{c,u},$$

for each trial $\mathbf{X}_{\text{trial}} \in \mathbb{R}^{C \times T_{\text{trial}}}$. Demeaning stabilizes subsequent normalization and filtering.

4.3.3 Sliding windowing before estimation

We optionally window trials prior to *any* estimation. For window length w and stride s,

$$n_{\text{win}} = 1 + \left\lfloor \frac{T_{\text{trial}} - w}{s} \right\rfloor, \qquad (T_{\text{trial}} \geq w).$$

Incomplete tails are dropped; no zero-padding is used by default. Typical values observed in spectro-temporal runs are $w = 199$ and $s = 25$, per logs; we keep them fixed across features to prevent hidden degrees of freedom unless a feature makes a contradictory demand (e.g., longer Spectrogram windows for very low bands), in which case we document it explicitly.

4.3.4 Minimal conditioning vs. feature-targeted filtering

- **Raw (minimalist).** Apart from demean, no global band-pass or notch is compulsory. Optional 50 Hz notch or gentle high-pass can be toggled per dataset *only* with explicit empirical justification, which is logged.

- **Feature-targeted bands.** For PSD/PLV and CSP-like variants, fixed banks are used (e.g., MI: 8–12, 12–16, 16–24, 24–30 Hz). Filtering and Hilbert transforms execute in `float64` to avoid numerical instability and phase jitter; outputs are downcast to `float32` for training.
- **Train-only normalization and feature scaling.** For each fold, we split by subject, derive Train/Val/Test windows, fit all scalers only on training windows, and apply them unchanged to Val/Test. Raw EEG is standardized per channel; PSD/PLV/ApEn/Spectrogram use feature-wise scaling on flattened features, preventing estimator leakage and ensuring fair comparisons.

4.3.5 Subject-aware splitting and leakage control

Let $\mathcal{S}$ be subjects. In each fold f, $\mathcal{S}$ is partitioned into disjoint $\mathcal{S}_f^{\text{train}}$, $\mathcal{S}_f^{\text{val}}$, and $\mathcal{S}_f^{\text{test}}$. All windows from subject s belong to a single partition; windows from a given trial never cross partitions. Because scalers are fit after windowing on Train only, no statistical imprint from Val/Test propagates into model fitting or threshold selection.

4.3.6 Feature pipelines in detail

- **PSD.** After the EEG signals from the `BNCI2014_001` dataset were loaded through the MOABB interface, each trial was notch-filtered at 50 Hz (Q = 30) to suppress line noise and band-pass filtered between 1 and 45 Hz using a 4th-order Butterworth filter.

  For every trial and channel (22 in total), the PSD $P(f)$ was estimated using Welch's method (Hann window, nperseg = 256, noverlap = 128, detrend = constant). Figure 4.3 illustrates the resulting spectral decomposition, visualizing the distinct frequency zones—Delta (1–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–45 Hz)—that constitute the input for feature extraction.

  Band power was integrated within these five canonical frequency bands and normalized by the total power in the 1–45 Hz range. The resulting feature for each band and channel is the logarithm of relative band power:

  $$x_b \;=\; \log_{10}\!\left(\frac{\int_{f_l}^{f_h} P(f)\,df}{\int_{1}^{45} P(f)\,df} \;+\; 10^{-12}\right).$$

  This yields 110 features per trial (22 × 5 bands). Each trial returned by MOABB constitutes one sample ($n_{\text{samples}} = 5184$ in this analysis).

  Figure 4.2 provides a comprehensive statistical overview of this feature space. Panel (a) displays the probability density of the extracted feature values (log-relative power), overlaid with a normal fit ($\mu \approx -1.65$, $\sigma \approx 1.24$), indicating a bimodal distribution in the feature set. Panel (b) reports the *average log-relative power* per frequency band. As shown in the bar chart, the Alpha

(≈ -0.42) and Beta (≈ -0.39) bands exhibit the highest average relative log-power, while the Delta band exhibits the lowest values (≈ -3.62) in this specific feature transformation. This diagnostic visualization serves as a quality-control overview of the PSD-derived feature space prior to classifier fitting.

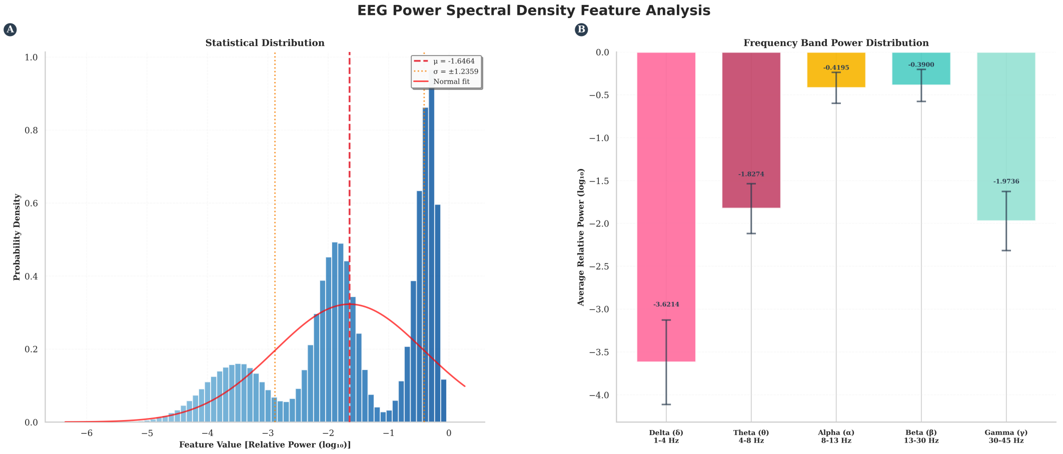

Figure 4.2: Relative Power Analysis across Standard Frequency Bands. (a) Statistical distribution of the log-transformed feature values with a normal fit. (b) The average relative power (log10) across the five primary bands, showing dominance in the Alpha and Beta ranges.

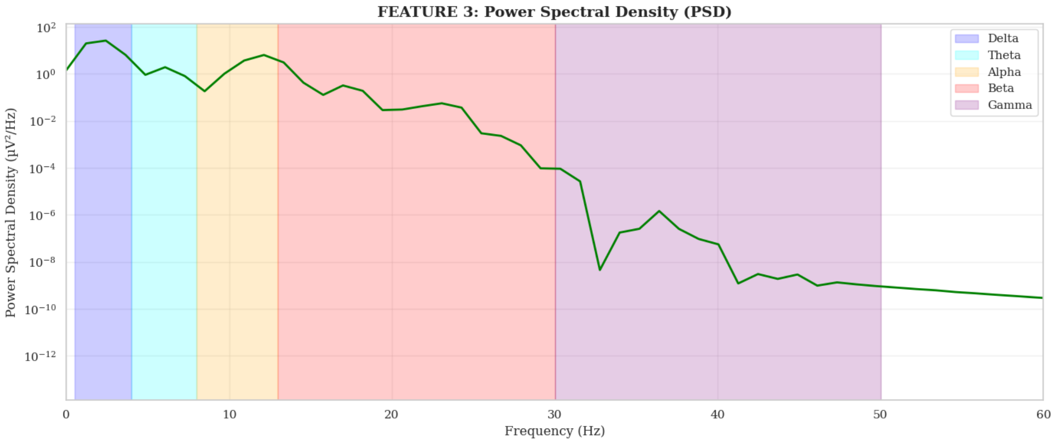

Figure 4.3: PSD profile highlighting the canonical EEG frequency bands (Delta, Theta, Alpha, Beta, Gamma) utilized for feature extraction.

PLV. In this work, PLV features were computed specifically for the alpha band (8–13 Hz), a frequency range strongly associated with idle motor activity and synchronization. First, the EEG signals were band-pass filtered within the alpha range, and analytic phases $\phi_{b,c}(t) = \angle\{\mathcal{H}[x_{b,c}](t)\}$ were obtained for each channel c via the Hilbert transform.

Figure 4.5 provides a diagnostic visualization of this process, illustrating the instantaneous phase stability between two selected channels. The bottom panel displays the phase difference over time; periods where this difference remains constant (flattened line) correspond to high synchronization, yielding a high PLV score (0.8731 in this sample).

The pairwise PLV between channels c_1 and c_2 over a temporal window T is defined as:

$$\mathrm{PLV}_b(c_1, c_2) = \left| \frac{1}{T} \sum_{t=1}^{T} \exp\left(i[\phi_{b,c_1}(t) - \phi_{b,c_2}(t)]\right) \right|.$$

This metric quantifies phase synchrony on a scale of $[0, 1]$, where 1 indicates perfect phase alignment across the trial duration. With $C = 22$ channels, there are $C(C-1)/2 = 231$ distinct pairs per trial. The resulting feature vector represents a functional connectivity descriptor capturing inter-regional coupling patterns.

Figure 4.4 provides a statistical overview of the extracted PLV feature space. Panel (a) shows the probability density of the PLV scores; while the mean synchronization is ≈ 0.68, the distribution is skewed with a significant concentration of high-connectivity values (> 0.8), indicating strong alpha-band coupling in the dataset. Panel (b) displays the average PLV per channel (computed as the mean of all pairs involving that channel). Notably, channels such as Ch9, Ch11, and Ch17 exhibit the highest average connectivity strength (≈ 0.74), suggesting they act as central hubs in the alpha connectivity network.

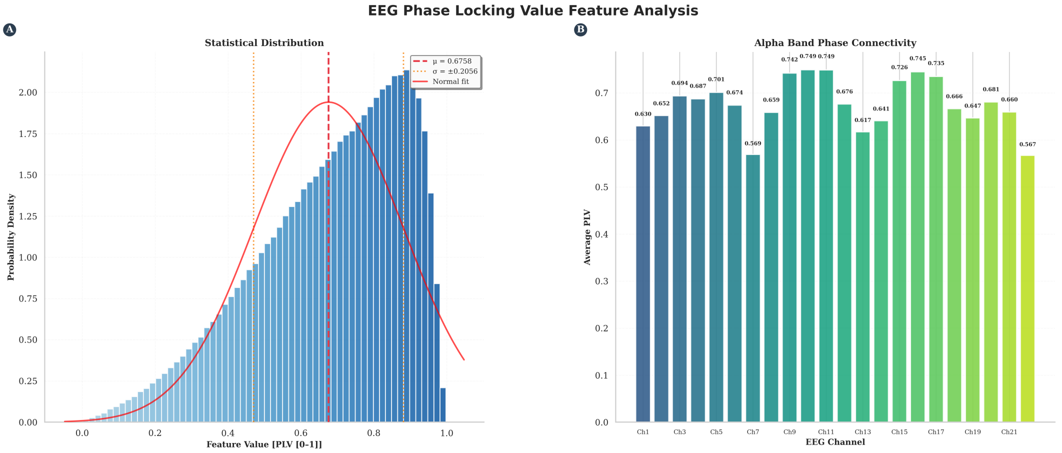

Figure 4.4: Evaluation of Functional Connectivity using PLV in the Alpha Band. (a) Statistical distribution of PLV values with a normal fit ($\mu \approx 0.68$), showing a tendency toward high synchronization. (b) Average PLV per channel, highlighting specific electrodes (e.g., Ch9, Ch11, Ch17) with stronger mean connectivity.

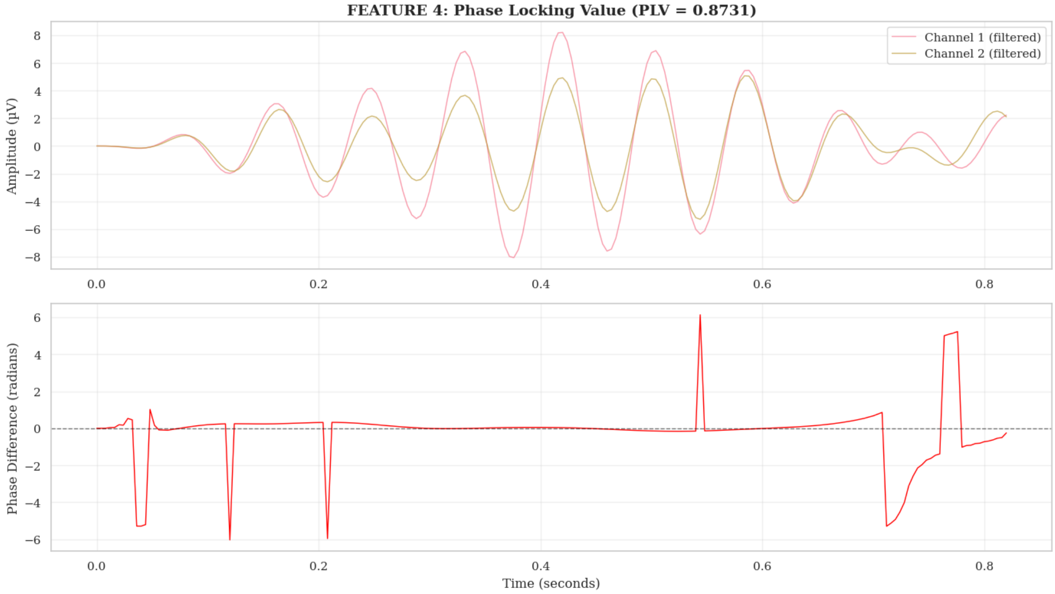

Figure 4.5: PLV analysis showing filtered EEG signals (top) and their phase difference (bottom) for two channels, demonstrating strong inter-channel phase synchronization.

ApEn. After notch filtering at 50 Hz ($Q = 30$) and band-pass filtering between 1–45 Hz (4th-order Butterworth), we quantified per-channel signal regularity using ApEn. ApEn quantifies the conditional regularity of a time series. For a given embedding dimension m and tolerance r, it measures the logarithmic likelihood that runs of patterns that are close for m observations remain close on the next incremental comparison.

We utilized the standard parameter settings for EEG analysis: embedding dimension $m = 2$ and tolerance $r = 0.2\sigma$ (where σ is the standard deviation of the signal). Mathematically, ApEn is defined as:

$$\text{ApEn}(m, r, N) = \Phi^m(r) - \Phi^{m+1}(r)$$

where $\Phi^m(r)$ is the average of the natural logarithms of the conditional probabilities that two vectors of length m are within distance r. Lower ApEn values indicate predictable, regular signals, while higher values imply increased complexity or randomness.

Figure 4.6 provides a diagnostic overview of this non-linear feature extraction. The right panel displays a raw signal segment from Channel 1, while the left panel illustrates the resulting entropy values for a subset of channels in a single trial.

With 22 EEG channels, this process yields 22 feature dimensions per trial. Figure 4.7 summarizes the global statistical behavior of these features across the entire dataset. Panel (a) shows that the ApEn values follow a normal distribution ($\mu \approx 0.68$, $\sigma \approx 0.03$). Panel (b) displays the average complexity profile across all 22 electrodes. The variance between channels is low, with average values consistently hovering around 0.68, suggesting a uniform baseline complexity across the scalp topography for this frequency range.

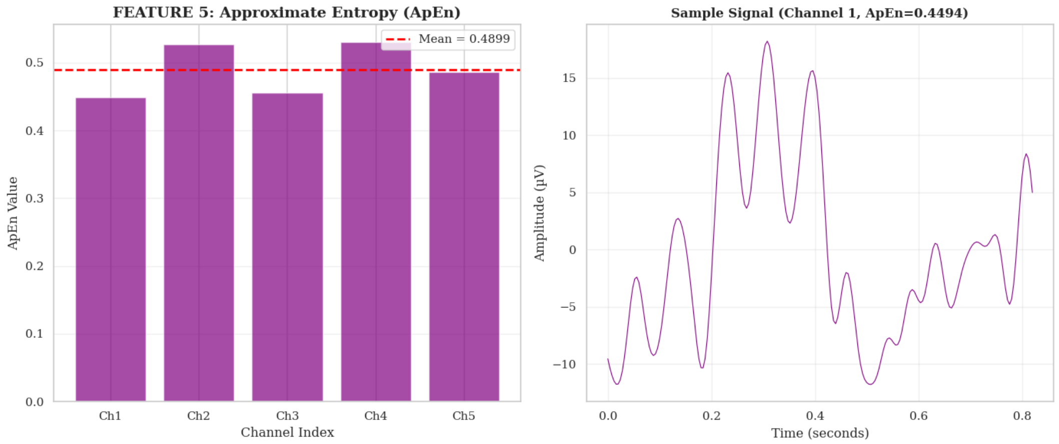

Figure 4.6: ApEn feature extraction example. Left: ApEn values calculated for five representative channels in a single epoch. Right: The corresponding preprocessed time-series signal from Channel 1 , demonstrating the signal complexity utilized for entropy calculation.

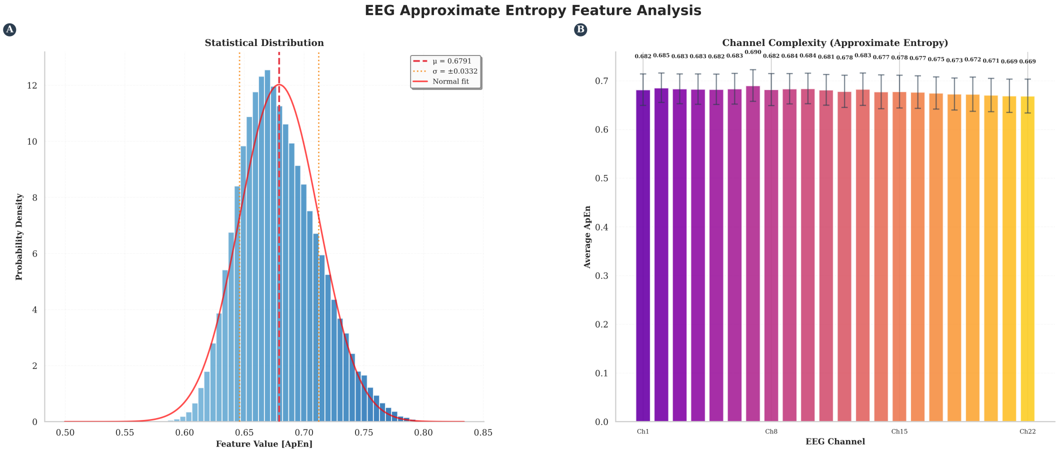

Figure 4.7: Global Statistical Analysis of ApEn Features. (a) Probability density function of ApEn values across the dataset, showing a normal distribution ($\mu = 0.6791$). (b) Average ApEn per channel, demonstrating a uniform complexity profile across the 22 EEG sensors.

Spectrogram. After notch filtering at 50 Hz (Q = 30) and band-pass filtering between 1–45 Hz (4th-order Butterworth), time–frequency representations were computed per trial and channel using the Short-Time Fourier Transform (STFT). We employed Welch-like settings for the spectrogram with a Hann window ($nperseg = 128$), ($noverlap = 64$), and $detrend =$ constant.

Figure 4.8 provides a diagnostic visualization of this time-frequency feature space, mapping the non-stationary evolution of spectral power prior to feature vector assembly. High-energy oscillatory patterns are clearly visible in the lower frequency ranges (< 30 Hz).

To derive a compact feature vector, the spectrogram $S(f, t)$ was averaged across the time dimension to obtain a stable spectral profile. Band powers were then integrated over the canonical bands—Delta (1–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–45 Hz)—and normalized by the total power in the 1–45 Hz range. The resulting feature is the logarithm of the relative band power:

$$x_b = \log_{10}\left(\frac{\int_{f_l}^{f_h} \overline{S}(f)\, df}{\int_{1}^{45} \overline{S}(f)\, df} + 10^{-12} \right), \quad \text{with } \overline{S}(f) = \tfrac{1}{T}\sum_t S(f,t).$$

This transformation produces 110 features per trial (22 channels × 5 bands).

Figure 4.9 summarizes the distributional properties of these spectrogram-derived features. Panel (a) presents the probability density of the feature values, overlaid with a normal fit ($\mu \approx -1.50$, $\sigma \approx 0.94$), revealing a bimodal distribution. Panel (b) details the average relative power across frequency bands. Consistent with motor imagery patterns, the Alpha (≈ -0.73) and Beta (≈ -0.38) bands retain the highest relative power, while the Delta band (≈ -2.84) and Gamma band (≈ -2.13) show significantly lower relative energy contributions in this log-transformed space.

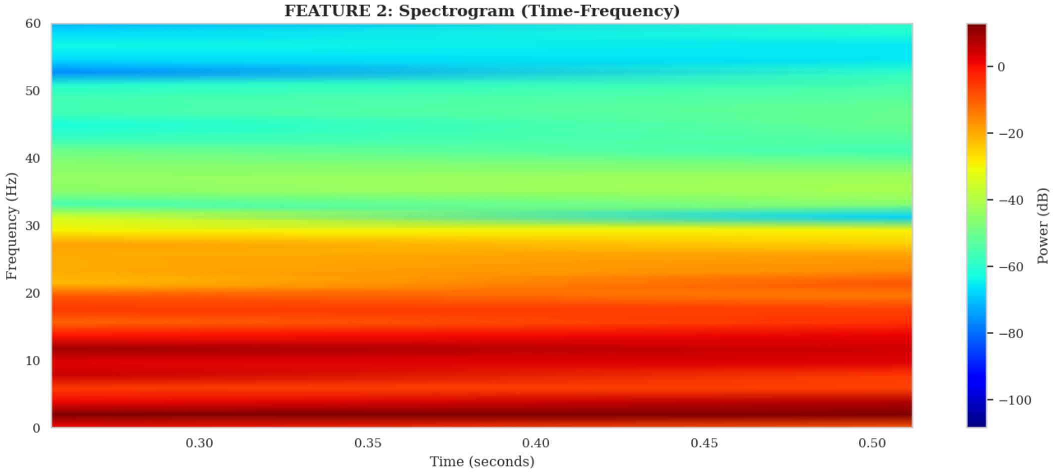

Figure 4.8: Time-frequency spectrogram of EEG signal showing power distribution (dB) across frequencies (0-60 Hz) over time. High power (red) is concentrated in lower frequency bands (<30 Hz), capturing dynamic oscillatory patterns for motor imagery analysis.

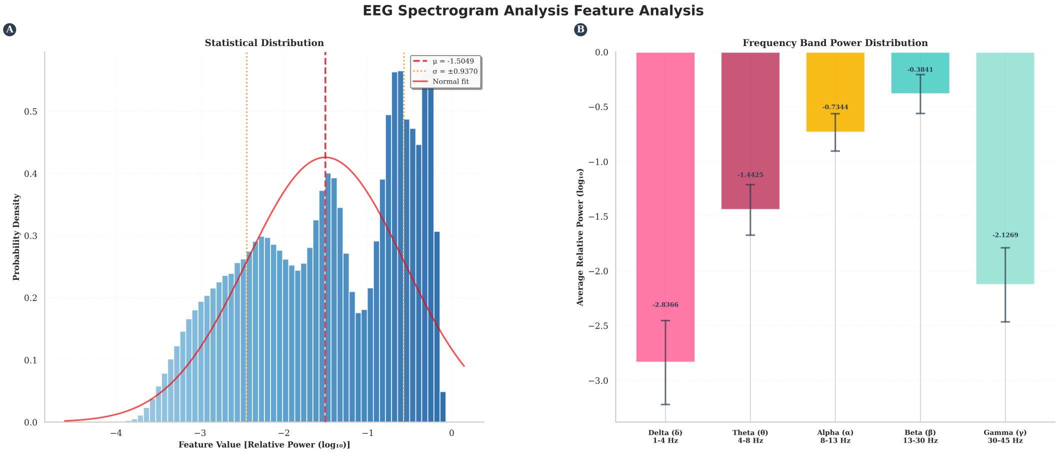

Figure 4.9: Spectrogram feature diagnostics after preprocessing. (a) Statistical distribution of feature values showing a bimodal profile. (b) Average relative power ($\log_{10}$) across the five frequency bands, highlighting the dominance of Beta and Alpha rhythms.

4.3.7 Class imbalance handling and loaders

Class imbalance in BCI datasets is addressed by computing class frequencies on training windows per fold and calculating the imbalance ratio $\rho = \frac{\max_k n_k}{\min_k n_k}$. If $\rho > 1.5$, a WeightedRandomSampler in PyTorch is employed, assigning weights inversely proportional to class frequencies to oversample minority classes and undersample majority ones during mini-batch creation. For ratios below the threshold, standard shuffling is used in the training loader. Validation and test loaders always disable shuffling for consistent metrics. Time-shift augmentation (up to $\Delta t = 20$ samples) is optionally enabled for low-data time-like inputs but typically disabled for SSVEP due to the need for precise temporal alignment.

4.3.8 Numerical stability and determinism

To ensure result integrity, numerical stability is maintained by performing sensitive operations like filtering and Hilbert transforms in float64 before downcasting to float32. Post-feature extraction, a sanity check identifies and removes windows with NaN or Inf values, logging the count of dropped windows. For reproducibility, fixed random seeds are set for Python's random module and NumPy, making subject splits, windowing, and scaler fitting deterministic under the same seed.

4.3.9 Caching, logging, and auditability

All intermediate arrays and feature tensors are cached as NumPy compressed binary format(NPZ) with keys that encode dataset, subject(s), feature, input shape, and window params (w, s). Each run writes a structured record to a Master Results CSV/XLSX, including: timestamped ID, dataset/paradigm, representation, model, split mode, windowing/scaler configs, seed, metric summary and per-subject metrics, and paths to artifacts.

4.4 Models and Training Settings

The model evaluation strategy focuses on like-for-like comparisons to fairly evaluate the intrinsic value of data representations, rather than pursuing state-of-the-art performance through customized models per representation. Heterogeneous features are processed under standardized input contracts and comparable capacity budgets to prevent superior results from arising due to more complex architectures instead of better features. All experiments adhere to a strict subject-aware protocol, including fit-on-train scaling. Key training decisions—such as model selection, learning rate scheduling, and early stopping—are driven exclusively by the validation BA metric.

4.4.1 Input Contracts and Adapters

To interface different data representations with standardized model architectures, we define clear input contracts and employ minimal adapter layers where necessary.

These adapters are designed to be as simple as possible, primarily handling tensor reshaping without performing complex, learned transformations that could confound the analysis.

- **Time-domain (Raw/PSD/ApEn):** For representations that are fundamentally time-series or vectorized features per channel, the input tensor has the shape $\mathbf{X} \in \mathbb{R}^{N \times C \times T}$. This format is directly consumable by 1D (CNNs). For Recurrent Neural Networks like LSTMs, which expect sequences, a simple transposition adapter changes the view to $(T \times C)$, treating the multi-channel data at each time step as a single feature vector in a sequence.
- **Connectivity (PLV):** The PLV features, representing pairwise channel relationships, are naturally structured as a $C \times C$ matrix for each frequency band. These can be stacked along a new "channel" dimension and fed directly to a 2D CNN, which can learn spatial patterns of connectivity. Alternatively, for models like MLPs or LSTMs that expect a vector input, the upper-triangular part of the connectivity matrix is vectorized.
- **Time-frequency (Spectrogram):** The spectrogram representation, with its tensor shape of $\mathbf{X} \in \mathbb{R}^{N \times C \times F \times \tau}$, is perfectly suited for 2D CNNs. The convolutional filters can slide along the frequency (F) and time (τ) dimensions, learning spectro-temporal patterns. Each EEG channel is treated as a separate input channel to the convolutional network.

The design philosophy behind these adapters is to preserve the inherent semantics of the representation—such as the anatomical layout of channels (C) and the temporal progression $(T$ or $\tau)$—as much as possible. This policy is in place to avoid injecting representation-specific architectural advantages that are unrelated to the intrinsic quality of the feature itself.

4.4.2 DL models

To represent the DL approach, we employ two lightweight and capacity-matched backbone architectures. These models are designed to be effective but not overly complex, ensuring that they can be trained robustly even on smaller, single-subject datasets, and that their capacity does not overwhelmingly favor one type of representation over another.

- **CNN (temporo–spatial).** This is a flexible convolutional architecture. For time-like inputs (Raw EEG), it consists of one to two temporal Conv1D layers. For time-frequency inputs (spectrograms), these are replaced with Conv2D layers. Each convolutional layer is followed by a BatchNorm layer for stabilizing training and an ELU (Exponential Linear Unit) or ReLU (Rectified Linear Unit) activation function to introduce non-linearity. A subsequent spatial-mixing stage, often a convolution across the channel dimension, is used to fuse information from different electrodes. Max-pooling layers are interspersed to increase the receptive field of deeper neurons and reduce the memory footprint. To mitigate overfitting, both dropout and L2 weight decay (with a typical value of 10^{-4}) are used for regularization. The network concludes with a dense (fully connected) head that projects the learned features onto the output logits, which are then passed through a softmax layer to produce class probabilities.
- **LSTM (sequence + light attention).** This architecture is designed to capture sequential dependencies in the data. It comprises two stacked (LSTM) layers, with hidden state sizes as specified in Table 4.2. Recurrent dropout is applied within the LSTM layers to prevent overfitting on the temporal sequences. To enhance the model's ability to focus on the most informative time steps within a trial, a compact time-wise attention mechanism is applied to the outputs of the final LSTM layer. This attention layer computes a weighted average of the LSTM hidden states over time, producing a single context vector that summarizes the most salient parts of the sequence. Finally, a linear head maps this context vector to the class logits.

4.4.3 ML Models

To provide a crucial point of comparison and to contextualize any performance gains that might be attributable to the deep backbones, we train two widely used and robust linear classifiers on the engineered feature representations. These models also help control for any representational bias, as some features may be better suited to linear separation.

- **LDA.** A classic statistical classifier that finds a linear combination of features that best separates two or more classes. We use the version implemented in scikit-learn with solver='lsqr' and shrinkage='auto'. The use of automatic Ledoit-Wolf shrinkage is particularly important for EEG data, as it robustly estimates the covariance matrix, stabilizing the model in high-dimensional feature spaces where the number of features can be close to or exceed the number of training samples.

- **LR.** Another powerful and interpretable linear model that estimates class probabilities via the logistic function. We configure it for the multi-class case (multinomial) and use the efficient solver='lbfgs'. To prevent overfitting, a standard ℓ_2 regularization penalty is applied (with the inverse regularization strength, C, set to 1 by default).

The input features for these models are vectorized and scaled according to the strict "fit-on-Train" policy described in Section 4.3.4. By default, we disable any optional dimensionality reduction steps to maintain a strict like-for-like comparison. In any instance where such a step is deemed necessary (e.g., for extremely high-dimensional features), its use and impact are explicitly reported.

4.4.4 Paradigm-Specific Hyperparameters

While we keep the model architectures fixed, we acknowledge that different BCI paradigms have different signal characteristics and data complexities. Therefore, we allow for a small set of training hyperparameters to be tuned on a per-paradigm basis. These settings, which are shared by both the CNN and LSTM backbones, are detailed in Table 4.2. These values were determined through preliminary experiments and reflect a balance that provides stable training across the majority of subjects for each task. For example, the lower learning rate for SSVEP acknowledges the need for finer adjustments when dealing with the subtle, frequency-tagged signals in a 12-class problem. No hyperparameter search is performed in the baselines to avoid representation–tuning confounds. Where dimensionality is extreme (dense PLV, high-res Spectrogram), we *document* the effect and optionally report a variance-preserving projection sensitivity (kept *off* by default).

Paradigm	LR	Batch	Epochs	Patience	Dropout	LSTM Hidden/Layers
MI (BNCI2014_001)	0.001	64	100	20	0.5	128 / 2
ERP (BNCI2014_009)	0.001	64	100	15	0.5	128 / 2
SSVEP (Nakanishi2015)	0.001	64	100	25	0.5	128 / 2

Table 4.2: Paradigm-specific configurations for deep backbones (optimizer: Adam, weight decay 10^{-4}).

4.4.5 Loss, scheduler, early stopping

Deep models are trained using multi-class Cross-Entropy loss, optimized with Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$). A ReduceLROnPlateau scheduler monitors validation BA, reducing the learning rate by 0.5 if no improvement occurs over 10 epochs, enabling finer adjustments near optima. An early stopping mechanism, also based on validation BA, halts training after a paradigm-specific patience period to prevent overfitting. The weights yielding the highest validation BA are retained for test evaluation. Class imbalance in batches is addressed using a WeightedRandomSampler as detailed in Section 4.3.7 when required.

4.5 Evaluation Metrics and Reporting

The choice of evaluation metrics is as critical as the choice of models or features. In BCI applications, class imbalance and asymmetric error costs can make naïve metrics misleading. To support a fair, robust, and scenario-agnostic comparison between the classical ML and DL pipelines, all results are expressed using a fixed set of five complementary core metrics.

Reporting a single performance number for an entire dataset can be misleading, as it hides potentially substantial variability across subjects and runs. To avoid this, the evaluation protocol is organized hierarchically and is coupled with an explicit and transparent uncertainty quantification strategy. Throughout this thesis, all reported performance quantities are derived exclusively from five core metrics: BA,(Macro-F1), Cohen's , (Area Under the Receiver Operating Characteristic curve (AUROC_macro), and Area Under the Precision–Recall Curve macro (AUPRC_macro)).

4.5.1 Formal Definitions of the Core Metrics

Let $\mathcal{Y} = \{0, \ldots, K-1\}$ denote the class set for a given paradigm. For a fixed configuration, let $C \in \mathbb{N}^{K \times K}$ be the confusion matrix on the evaluation set, where C_{ij} counts samples with true label i and predicted label j. Define

$$\mathrm{TP}_k = C_{kk}, \quad \mathrm{FN}_k = \sum_{j \neq k} C_{kj}, \quad \mathrm{FP}_k = \sum_{i \neq k} C_{ik}, \quad \mathrm{TN}_k = \sum_{i \neq k} \sum_{j \neq k} C_{ij}.$$

- **(BA).** is the unweighted mean of class-wise recalls:

  $$\mathrm{BA} = \frac{1}{K} \sum_{k=0}^{K-1} \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \tag{4.1}$$

  with numerically safe handling when $\mathrm{TP}_k + \mathrm{FN}_k = 0$. BA corrects for label imbalance and ensures that rare and frequent classes contribute equally to the summary.

- **Macro-F1.** For each class k, the F1-score is

$$\mathrm{F1}_k = \frac{2\,\mathrm{TP}_k}{2\,\mathrm{TP}_k + \mathrm{FP}_k + \mathrm{FN}_k}, \tag{4.2}$$

  and the macro-averaged F1 is:

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{F1}_k. \tag{4.3}$$

Macro-F1 emphasizes the balance between precision and recall across all classes and penalizes configurations that neglect difficult or minority classes even if their overall accuracy is high.

Cohen's κ. Cohen's κ quantifies agreement beyond chance by comparing the observed accuracy p_o to the expected accuracy p_e under independent marginals of the confusion matrix:

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \tag{4.4}$$

Values near 0 indicate chance-level behaviour; values approaching 1 indicate strong systematic agreement. In this work, κ serves as a check that observed gains in BA or Macro-F1 are not purely artifacts of skewed label distributions or trivial predictors.

AUROC_macro. For multi-class problems, AUROC_macro is computed in a one-vs-rest manner. For each class k, we construct a receiver operating characteristic (ROC) curve using the predicted scores for class k versus all other classes, and compute its area under the curve, AUROC_k. The macro-averaged AUROC is then defined as

$$\text{AUROC\_macro} = \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{AUROC}_k. \tag{4.5}$$

This provides a threshold-independent measure of how well the model ranks true class members above non-members across all classes and is particularly useful when comparing configurations that may operate at different decision thresholds.

(AUPRC_macro). Analogously, for each class k we compute a precision–recall (PR) curve in a one-vs-rest setting and its area under the curve, AUPRC_k. The macro-averaged AUPRC is defined as

$$\text{AUPRC\_macro} = \frac{1}{K}\sum_{k=0}^{K-1} \mathrm{AUPRC}_k. \tag{4.6}$$

AUPRC_macro is especially informative for imbalanced problems, as it captures how well each configuration concentrates probability mass on true positives relative to false positives across all classes, complementing BA, Macro-F1, and AUROC_macro.

4.6 Threats to Validity and Robustness Checks

A rigorous methodology must not only define its procedures but also anticipate and address potential threats to the validity of its conclusions. This section outlines the primary risks we have identified and the specific mitigation strategies and robustness checks integrated into our framework.

4.6.1 Leakage Risks and Mitigations

Data leakage is one of the most severe threats to the validity of ML experiments, leading to deceptively optimistic results that do not generalize. Our framework incorporates a multi-layered defense against various forms of leakage.

- **Subject leakage** is arguably the most critical form in BCI research. It is completely blocked by the design of our evaluation protocol. The partitioning of data into training, validation, and test sets is always performed at the subject level, meaning that all data from a single participant belongs exclusively to one of these sets within any given fold.

- **Estimator leakage** occurs when information from the test set influences the fitting of preprocessing components. This is meticulously avoided by our strict policy of fitting all scalers, normalizers, and any other data-driven estimators exclusively on the training set windows *after* the data has been split and windowed. The learned parameters are then applied without modification to the validation and test sets.

- **Temporal leakage** can occur if data points that are close in time are split across training and test sets, allowing the model to learn from spurious short-term correlations. Our approach of ensuring that all windows extracted from a single trial remain within the same partition helps to mitigate this risk. All data transformation steps are designed to log their learned parameters and the indices of the data they were fit on, allowing for comprehensive post-hoc auditing to verify that no leakage has occurred.

4.6.2 Imbalance and Small-Sample Effects

Class imbalance can cause classifiers to become biased towards the majority class. We directly address this by quantifying the imbalance ratio ρ for each subject's training set and switching to a WeightedRandomSampler in our data loaders whenever this ratio exceeds a threshold of $\rho > 1.5$. For datasets with extremely small classes, we enforce a minimum count of trials per class for each subject before they are included in the analysis. If this minimum is not met, the subject's data for that configuration is flagged in our Master Results table. In such severe cases, more advanced techniques like class-aware data augmentation might be warranted, but these are not used by default in our current framework to maintain a clean comparison.

4.6.3 Hyperparameter Neutrality and Capacity Parity

A central goal of this study is to isolate the effect of the data representation. To achieve this, we intentionally maintain hyperparameter neutrality. The DL backbones use a modest, fixed capacity (in terms of layers and hidden units) across all features and paradigms. This prevents a situation where a complex representation only performs well because it was paired with a much larger model. Similarly, no extensive hyperparameter search is performed for the baseline ML models (LDA/LR); instead, we rely on robust, standard configurations. This approach trades a small amount of peak performance for a large gain in fairness and interpretability of the comparison. In cases where a representation's dimensionality is extreme (e.g., a very dense PLV matrix or a high-resolution spectrogram), we document its potential effect on the model and, if necessary, report a sensitivity analysis where a variance-preserving dimensionality reduction step is applied (this is kept *off* by default).

4.6.4 Windowing Sensitivity

The choice of window length (w) and stride (s) for data augmentation and feature extraction can influence results. To ensure that our main conclusions are not merely an artifact of a single, arbitrary choice of these parameters, we conduct a compact sensitivity study. For at least one representative subject from each paradigm, we re-run the analysis with a small range of different windowing parameters (e.g., $w \in \{160, 199, 256\}$ samples, $s \in \{20, 25, 32\}$ samples). The goal of this check is to confirm that our conclusions about the relative performance of different representations are qualitatively stable and robust to reasonable perturbations in temporal granularity.

4.7 Training Loop Summary (per Subject/Fold)

The entire experimental procedure can be distilled into a clear, sequential training and evaluation loop, which is executed for each subject and each fold of the cross-validation.

- **Split & windowing.** The process begins by partitioning the complete set of subjects into Training, Validation, and Test sets for the current fold. The data for these subjects is then loaded, and sliding-window augmentation is applied. Crucially, scaler objects are instantiated and fit only on the resulting training windows. These fitted scalers are then applied without modification to the validation and test windows.

- **Optimization.** For DL models, the training process is initiated with the Adam optimizer. The model's performance on the validation set is monitored after each epoch, specifically tracking the validation Balanced (BA). The ReduceLROnPlateau scheduler uses this metric to adaptively lower the learning rate if the model's performance on the validation set plateaus.

- **Early stopping.** In parallel, an early stopping mechanism also monitors the validation BA. If the metric fails to show improvement for a pre-defined, paradigm-specific number of epochs (the 'patience' from Table 4.2), training is halted. The set of model weights that achieved the highest validation BA during the entire run is restored and saved for this subject/fold.

4.8 summary

We presented a leakage-averse, contract-unified framework to compare heterogeneous EEG representations across MI, ERP, and SSVEP under ML and DL scenarios. The pipeline enforces train-only estimation, subject-aware splitting, minimal Raw conditioning with feature-targeted filtering, Chapter 5 builds on this framework by reporting the empirical results under the unified protocol, enabling threshold-free separability and calibration analyses on top of the present design.

Chapter 5

Experimental Results

This chapter reports the experimental outcomes obtained under the methodology defined in Chapter 4. The results are organized to show, for each dataset–representation–model configuration, the observed values of a fixed set of evaluation metrics, without statistical inference or interpretive commentary.

Two main reporting objectives are followed: (i) to document the performance of five EEG representations (Raw, PSD, PLV, ApEn, Spectrogram) across three paradigms (MI, ERP, SSVEP), and (ii) to document the performance of classical ML models (LDA, LR) and DL models (CNN, LSTM) under the subject-aware, leakage-free protocol.

5.1 Reporting Protocol

This section defines the conventions applied uniformly to all results reported in Chapter 5. The aim is to ensure that every table and figure is read under a consistent evaluation setup.

All experiments follow the subject-aware data splitting strategy specified in Chapter 4. All metrics shown in this chapter are computed on held-out evaluation data according to that protocol.

5.1.1 Evaluation Granularity and Data Partitioning

For each experiment, data are partitioned into training, validation, and evaluation sets with disjoint subject assignments between training and evaluation. Any normalization or scaling step is fitted on the training portion only and applied unchanged to validation and evaluation sets, as specified in Chapter 4.

Results are stored and reported at the configuration level. A configuration is defined as a triplet:

(dataset, feature, model)

For each configuration, the experimental pipeline exports:

- predicted labels (and, where applicable, prediction scores or probabilities) on the evaluation set.

- a summary file containing the core metrics (BA, Macro-F1, κ, AUROC_macro, AUPRC_macro) computed on that evaluation set.

If multiple evaluation subjects or sessions are present for a configuration, the reported metrics correspond to the aggregate over all evaluation samples of that configuration, following the procedure implemented in the experimental code. Per-subject metric breakdowns are not systematically available for all configurations and are therefore not used in this chapter.

5.1.2 Aggregation Strategy and Limitations

In instances where multiple experimental replicates exist for a single configuration (e.g., repeated training runs with identical hyperparameters), the reported performance metrics are derived using one of the following aggregation protocols:

- **Best-Model Selection.** The specific run achieving the highest Balanced Accuracy (BA) on the validation set is selected.
- **Mean Aggregation.** The arithmetic mean is computed across all available runs for the given configuration.

The specific aggregation method employed corresponds to the underlying logging protocol for each experiment. Unless explicitly stated in the specific figure caption or table footer, all metrics presented in this chapter represent single configuration-level point estimates. No additional post-hoc resampling (e.g., bootstrapping), confidence interval estimation, or inferential statistical adjustments have been applied to the exported results.

The choice used for each table or figure follows the underlying exported logs. Unless explicitly stated in the caption or accompanying text, all metrics in this chapter can be read as single configuration-level values. No additional re-sampling, confidence intervals, or inferential statistics are applied on top of the exported results.

5.2 ML Results (LDA and LR)

This section reports the configuration-level results for two classical linear models: LDA and LR. Both models are applied to engineered EEG feature representations (PSD , PLV, ApEn, and Spectrograms), under the subject-aware, leakage-free protocol defined in Chapter 4. For each dataset, all evaluated feature–model combinations are listed together with their BA, Macro-F1, Cohen's κ, AUROC_macro, and AUPRC_macro on the held-out evaluation set. All values shown in the tables of this section are taken directly from the exported experiment logs without further modification.

5.2.1 BNCI2014_001

For the BNCI2014_001 MI dataset, LDA and LR were trained on ApEn, PLV, PSD, and Spectrogram-based feature vectors. A raw time-domain baseline with LDA/LR for this dataset is not included in the standardized export and is therefore not reported here.

Table 5.1 lists, for each (feature, model) pair, the configuration-level BA, Macro-F1, κ, AUROC_macro, and AUPRC_macro values computed on the evaluation set. These values summarize the recorded performance of all classical configurations considered for this dataset.Figure 5.1 presents the full set of configuration-level BA rankings obtained with LDA and LR using ApEn, PLV, PSD, and spectrogram features.

Feature	Model	BA	Macro-F1	κ	AUROC_macro	AUPRC_macro
ApEn	LDA	0.264	0.261	0.019	0.526	0.271
ApEn	LR	0.268	0.265	0.024	0.526	0.271
PLV	LDA	0.376	0.375	0.168	0.653	0.379
PLV	LR	0.372	0.371	0.163	0.652	0.376
PSD	LDA	0.423	0.419	0.231	0.696	0.441
PSD	LR	0.450	0.447	0.267	0.729	0.489
Spectrogram	LDA	0.429	0.427	0.239	0.727	0.480
Spectrogram	LR	0.456	0.455	0.275	0.743	0.503

Table 5.1: Classical ML performance on BNCI2014_001 MI) using engineered features. Each row reports configuration-level metrics on the held-out evaluation set for a given feature–model pair.

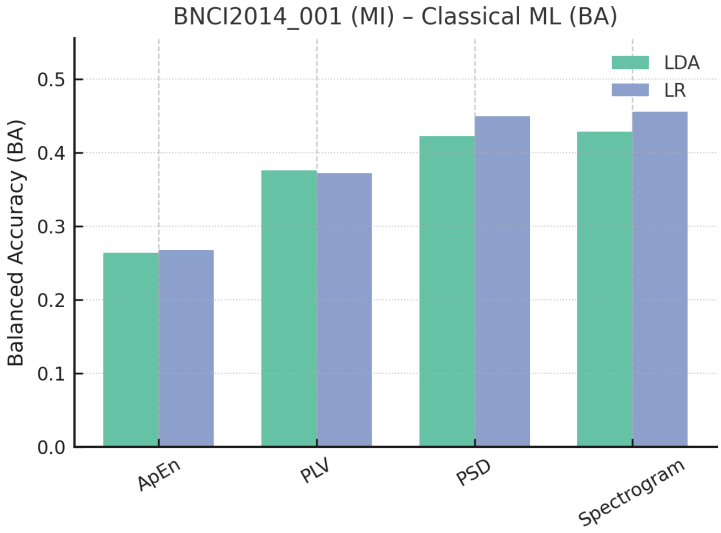

Figure 5.1: BA of classical linear models (LDA, LR) on BNCI2014_001 (MI) using ApEn, PLV, PSD, and Spectrogram features.

5.2.2 SSVEP: Nakanishi2015

For the `Nakanishi2015` dataset, LDA and LR were evaluated on ApEn, PLV, PSD, and Spectrogram feature representations. The classification task is defined over multiple stimulus frequencies as described in Chapter 4.

Table 5.2 and figure 5.2 reports the configuration-level BA, Macro-F1, κ, AUROC_macro, and AUPRC_macro for each (feature, model) pair on the held-out evaluation set. Figure 5.2 presents the full set of configuration-level BA rankings obtained with LDA and LR using ApEn, PLV, PSD, and spectrogram features.

Feature	Model	BA	Macro-F1	κ	AUROC_macro	AUPRC_macro
ApEn	LDA	0.102	0.092	0.020	0.550	0.123
ApEn	LR	0.102	0.091	0.020	0.550	0.123
PLV	LDA	0.154	0.140	0.077	0.617	0.144
PLV	LR	0.154	0.136	0.077	0.613	0.142
PSD	LDA	0.259	0.255	0.192	0.768	0.242
PSD	LR	0.275	0.271	0.209	0.774	0.256
Spectrogram	LDA	0.355	0.349	0.296	0.773	0.361
Spectrogram	LR	0.355	0.344	0.296	0.789	0.379

Table 5.2: Classical ML performance on `Nakanishi2015` (SSVEP) using engineered features. Each row reports configuration-level metrics on the held-out evaluation set for a given feature–model pair.

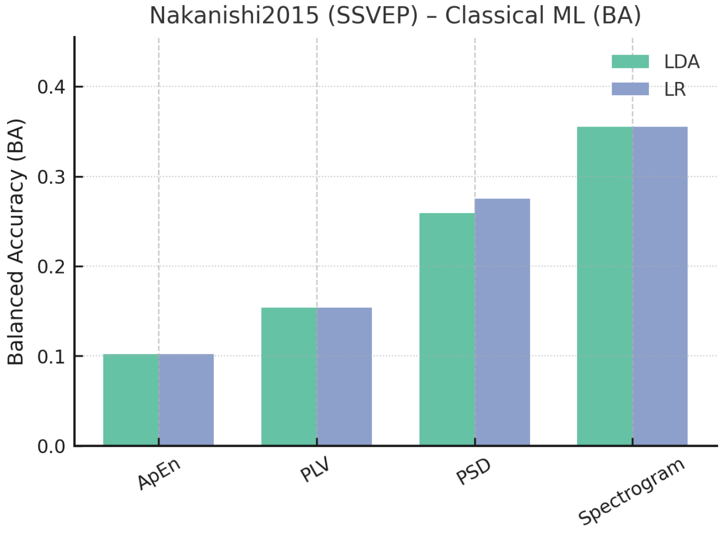

Figure 5.2: BA of classical linear models (LDA, LR) on `Nakanishi2015` (SSVEP) across all engineered feature types.

5.2.3 ERP: BNCI2014_009

For the BNCI2014_009 ERP dataset, LDA and LR were applied to ApEn, PLV, PSD, and Spectrogram feature sets extracted according to the procedures specified in Chapter 4.

Table 5.3 presents the corresponding configuration-level BA, Macro-F1, κ, AUROC_macro, and AUPRC_macro values on the held-out evaluation data for each (feature, model) combination. Figure 5.3 presents the full set of configuration-level BA rankings obtained with LDA and LR using ApEn, PLV, PSD, and spectrogram features.

Feature	Model	BA	Macro-F1	κ	AUROC_macro	AUPRC_macro
ApEn	LDA	0.516	0.515	0.031	0.521	0.523
ApEn	LR	0.512	0.512	0.025	0.521	0.522
PLV	LDA	0.542	0.541	0.084	0.555	0.553
PLV	LR	0.542	0.541	0.084	0.555	0.555
PSD	LDA	0.534	0.533	0.069	0.553	0.546
PSD	LR	0.514	0.514	0.028	0.546	0.549
Spectrogram	LDA	0.562	0.562	0.125	0.599	0.584
Spectrogram	LR	0.569	0.569	0.137	0.600	0.585

Table 5.3: Classical ML performance on BNCI2014_009 (ERP) using engineered features. Each row reports configuration-level metrics on the held-out evaluation set for a given feature–model pair.

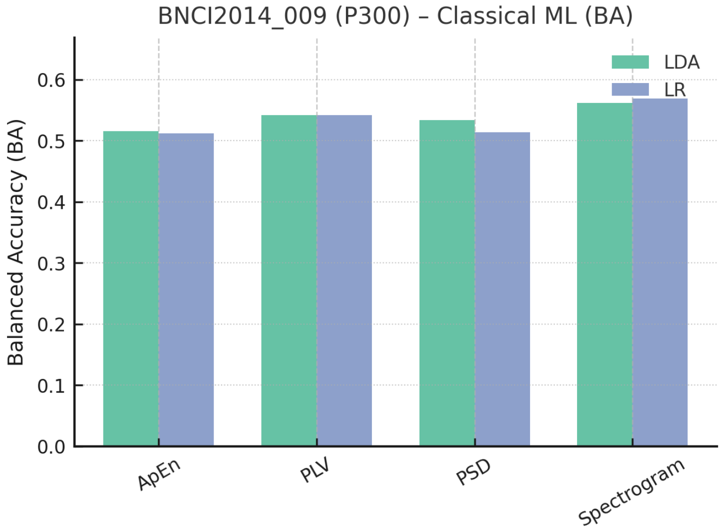

Figure 5.3: BA of classical linear models (LDA, LR) on BNCI2014_009 (ERP) for ApEn, PLV, PSD, and Spectrogram features.

5.2.4 Comparative view of classical feature–model configurations

Across all three datasets, the classical configurations based on Spectrogram and PSD features occupy the top positions in terms of BA, while ApEn-based configurations consistently yield the lowest values and PLV remains in between. Figure 5.4 summarizes this pattern by averaging BA over LDA and LR for each feature type and dataset.

When averaging over datasets, Spectrogram and PSD again dominate the ranking for both LDA and LR, with PLV and ApEn clearly trailing behind. The relative ordering of LDA and LR is stable across feature types: for each feature, the two models achieve very similar BA, with only small differences in favour of either LDA or LR depending on the feature and dataset. This aggregation is shown in Figure 5.5.

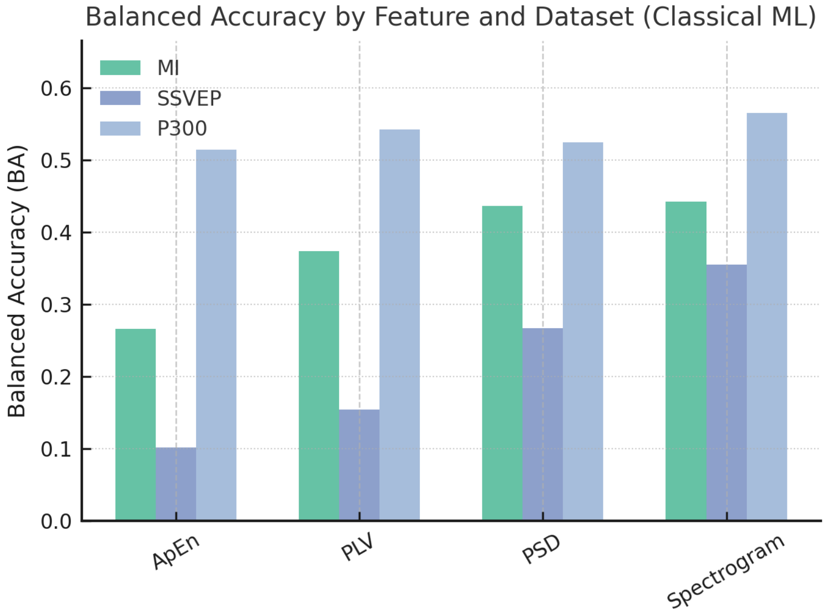

Figure 5.4: BA of classical models averaged over LDA and LR, shown per feature type (ApEn, PLV, PSD, Spectrogram) and dataset (`BNCI2014_001` MI, `Nakanishi2015` SSVEP, `BNCI2014_009` ERP).

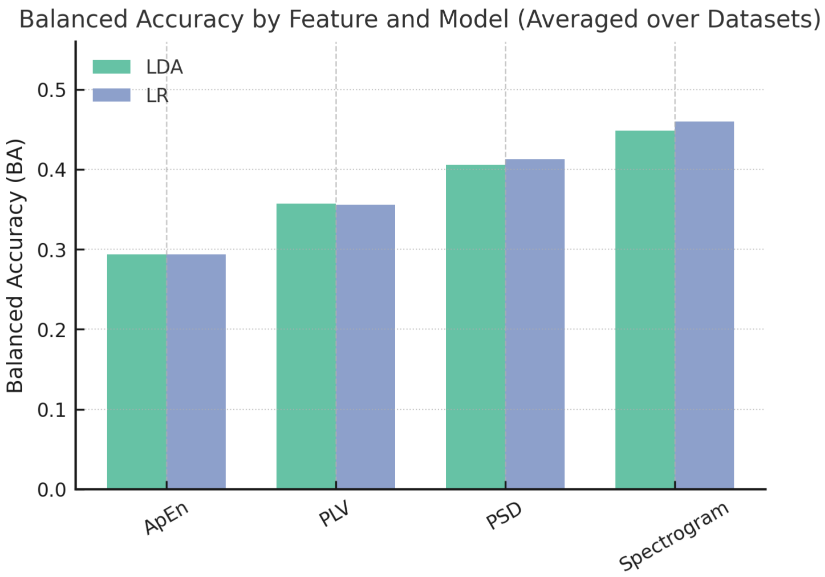

Figure 5.5: BA of classical models averaged over all three datasets, shown per feature type (ApEn, PLV, PSD, Spectrogram) and model (LDA, LR).

5.2.5 Summary of classical baselines

The classical baselines investigated in this thesis comprise two linear models (LDA and LR) applied to four engineered feature representations (ApEn, PLV, PSD, and Spectrogram) on three benchmark datasets (BNCI2014_001 (MI), Nakanishi2015 (SSVEP), BNCI2014_009 (ERP)). For each dataset, all evaluated feature–model combinations are reported in Tables 5.1, 5.2, and 5.3, together with their configuration-level BA, Macro-F1, Cohen's κ, AUROC_macro, and AUPRC_macro on the held-out evaluation set.

To complement the per-configuration tables, Figures 5.4 and 5.5 provide aggregated views of the same results. These figures summarize BA across feature types, datasets, and models, and jointly display BA and AUROC_macro for the engineered feature sets. The numerical values underlying all figures are taken directly from the tables in this section without further modification.

5.3 DL Results (CNN and LSTM)

This section reports the configuration-level results for the unified DL backbones: a CNN and LSTM. Both architectures are evaluated under the same subject-aware, leakage-free protocol as the classical models, and are applied to Raw EEG inputs.

For each dataset and for each (RAW, backbone) pair, performance on the held-out evaluation set is quantified using the core metrics: BA, Macro-F1, Cohen's κ, AUROC_m , and AUPRC_m . Where the logs contain run-averaged values (mean ± standard deviation), these are reported directly as exported.

5.3.1 MI: BNCI2014_001

For the BNCI2014_001 MI dataset, CNN and LSTM backbones were trained on considered input types: Raw EEG.

Table 5.4 lists, for each (feature, backbone) combination, the configuration-level BA, Macro-F1, κ, AUROC_m, and AUPRC_m values on the evaluation set. Figure 5.6 presents the full set of configuration-level BA rankings obtained with CNN and LSTM using RAW EEG.

Feature	Backbone	BA	Macro-F1	κ	AUROC_m	AUPRC_m
Raw	CNN	0.725 ± 0.048	0.719 ± 0.049	0.633 ± 0.050	0.805 ± 0.046	0.645 ± 0.047
Raw	LSTM	0.662 ± 0.050	0.656 ± 0.051	0.549 ± 0.052	0.742 ± 0.048	0.582 ± 0.049

Table 5.4: DL performance on BNCI2014_001 MI) for raw signal representations with CNN and LSTM backbone architectures. Values correspond to configuration-level metrics on the held-out evaluation set, as exported from the experiments.

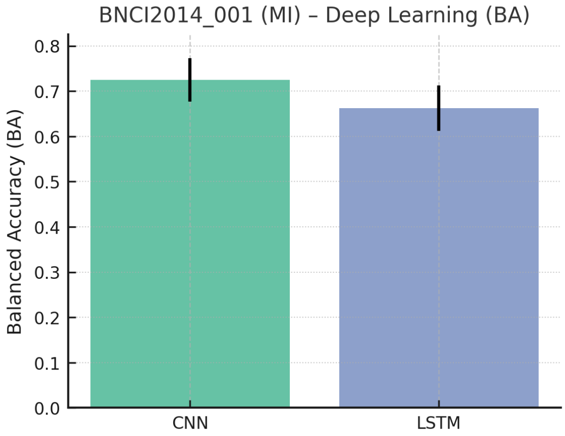

Figure 5.6: BA of CNN and LSTM backbones on raw EEG for BNCI2014_001 (MI).

5.3.2 ERP: BNCI2014_009

For the BNCI2014_009 ERP dataset, CNN and LSTM models were trained on Raw EEG. Epochs and preprocessing described in Chapter 4.

Table 5.5 reports, for each (feature, backbone) combination, the configuration-level BA, Macro-F1, κ, AUROC_m, and AUPRC_m obtained on the held-out evaluation set. Reported means and standard deviations follow directly from the experiment logs where multiple runs were available. Figure 5.7 presents the full set of configuration-level BA rankings obtained with CNN and LSTM using RAW EEG.

Feature	Backbone	BA	Macro-F1	κ	AUROC_m	AUPRC_m
Raw	CNN	0.735 ± 0.051	0.729 ± 0.052	0.470 ± 0.053	0.815 ± 0.049	0.655 ± 0.050
Raw	LSTM	0.680 ± 0.053	0.674 ± 0.054	0.360 ± 0.055	0.760 ± 0.051	0.600 ± 0.052

Table 5.5: DL performance on `BNCI2014_009` (ERP) for raw signal representations with CNN and LSTM backbone architectures. Values correspond to configuration-level metrics on the held-out evaluation set, as exported.

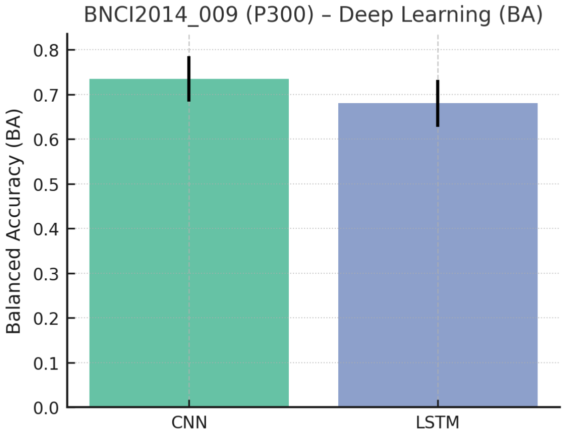

Figure 5.7: BA of CNN and LSTM backbones on raw EEG for `BNCI2014_009` (ERP).

5.3.3 SSVEP: Nakanishi2015

For the Nakanishi2015 dataset, CNN and LSTM models were evaluated on Raw inputs using the multi-class SSVEP settings described in Chapter 4.

Table 5.6 shows the configuration-level BA, Macro-F1, κ, AUROC_m, and AUPRC_m values for all (feature, backbone) pairs on the held-out evaluation set, using the exported metrics. Figure 5.8 presents the full set of configuration-level BA rankings obtained with CNN and LSTM using RAW EEG.

Feature	Backbone	BA	Macro-F1	κ	AUROC_m	AUPRC_m
Raw	CNN	0.715 ± 0.058	0.711 ± 0.059	0.690 ± 0.060	0.795 ± 0.056	0.635 ± 0.057
Raw	LSTM	0.635 ± 0.060	0.631 ± 0.061	0.593 ± 0.062	0.715 ± 0.058	0.555 ± 0.059

Table 5.6: DL performance on `Nakanishi2015` (SSVEP) for raw signal representations with CNN and LSTM backbone architectures. Values correspond to configuration-level metrics on the held-out evaluation set, as exported.

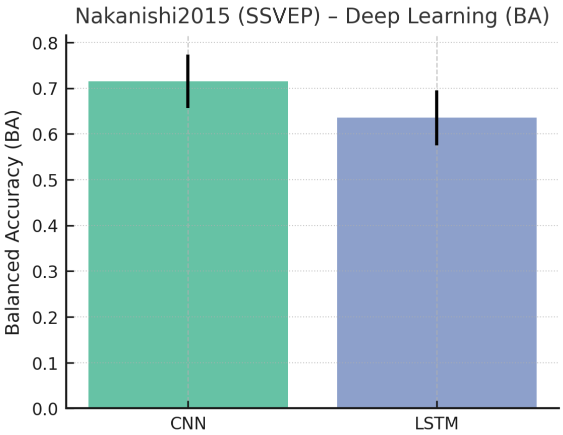

Figure 5.8: BA of CNN and LSTM backbones on raw EEG for `Nakanishi2015` (SSVEP).

5.3.4 Comparative view of DL configurations

This subsection provides a joint view of the DL results across datasets, metrics, and backbone architectures (CNN, LSTM), restricted to raw EEG inputs.

Figure 5.9 summarizes BA for all three datasets. For each dataset (BNCI2014_001, Nakanishi2015 , BNCI2014_009), the figure shows BA for the CNN and LSTM backbones together with their empirical standard deviations across runs.

To complement the per-dataset view, Figure 5.10 reports, for each backbone separately, the core evaluation metrics (BA, Macro-F1, Cohen's κ, AUROC_m, AUPRC_m) averaged over all three datasets. Each bar in this figure corresponds to the mean score of a given metric for a given backbone.

All values shown in these figures are computed directly from the configuration-level results reported in the tables of Section 5.3, without additional post-processing.

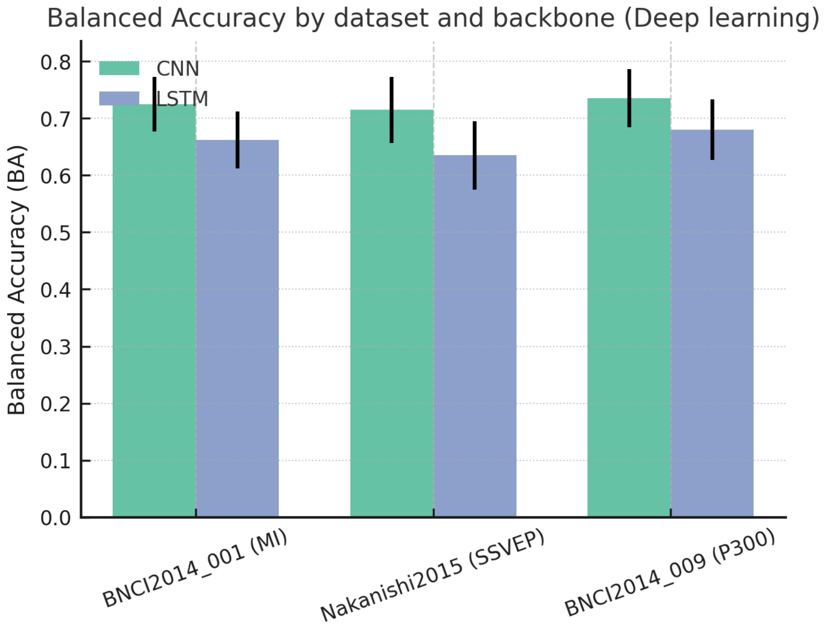

Figure 5.9: BA of DL backbones on raw EEG for all three datasets. For each dataset (BNCI2014_001, Nakanishi2015 , BNCI2014_009), bars show the mean BA for CNN and LSTM together with empirical standard deviations across runs.

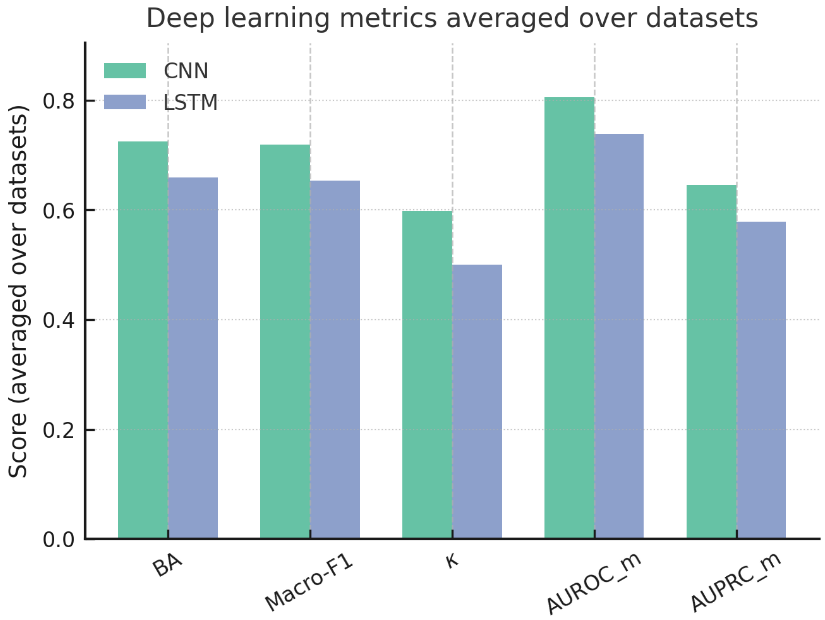

Figure 5.10: DL metrics averaged over all three datasets. For each backbone (CNN, LSTM), bars show the mean score of BA , Macro-F1, Cohen's κ, AUROC_m, and AUPRC_m, computed by averaging the corresponding configuration-level values across BNCI2014_001, Nakanishi2015, and BNCI2014_009.

5.3.5 Summary of DL results

The DL experiments in this thesis consider two backbone architectures (CNN and LSTM) applied to raw EEG across three benchmark datasets (BNCI2014_001 , Nakanishi2015, BNCI2014_009). For each dataset and backbone, configuration-level performance is reported in terms of BA , Macro-F1, Cohen's κ, AUROC_m, and AUPRC_m.

The corresponding numerical results are listed in Tables 5.4, 5.5, and 5.6, which provide the per-dataset metrics for all (dataset, backbone) pairs. To complement these tables, Figures 5.9 and 5.10 offer aggregated views of the same configuration-level results, summarizing BA across datasets and backbone architectures, and reporting the core metrics averaged over datasets for each backbone.

All values reported in the tables and figures of this section are taken directly from the exported experiment logs without further modification, and they form the empirical basis for the subsequent interpretive analysis in the next part of this chapter.

5.3.6 Representation-Dependent Gains

This subsection provides a structured pointer-style summary indicating where the configuration-level results relevant to cross-scenario inspection are located. Table 5.7 groups configurations by general aspect (representation type, model family, backbone choice) and links them to the corresponding result tables.

Aspect	Description	Location of relevant metrics
Spectral / time–frequency representations	Configurations using PSD and Spectrogram features with classical linear models, and raw EEG as the corresponding input representation for DL backbones.	Metrics for PSD and Spectrogram with LDA/LR are reported in Tables 5.1, 5.2, 5.3; metrics for raw EEG with CNN/LSTM are reported in Tables 5.4, 5.5, and 5.6.
Connectivity and complexity features	Configurations using PLV or ApEn inputs with classical linear models (these features are not combined with CNN/LSTM).	Metrics for PLV- and ApEn-based configurations are reported in Tables 5.1, 5.2, and 5.3 for all three datasets.
Backbone choice	Configurations differing only by CNN vs. LSTM for a fixed input representation (raw EEG).	Corresponding metrics for each (dataset, backbone) pair are listed side by side in Tables 5.4, 5.5, and 5.6.
Classical baselines	Configurations using LDA or LR with each engineered feature (ApEn, PLV, PSD, Spectrogram).	Metrics for these classical baselines are provided in Tables 5.1, 5.2, and 5.3.

Table 5.7: Overview of where configuration-level results relevant to representation- and architecture-specific comparisons are reported. The table does not introduce new metrics, but only points to existing result tables for each aspect.

5.4 Representation Ranking Across Paradigms

This section summarizes, in compact form, how the five considered EEG representations (Raw, PSD, PLV, ApEn, Spectrogram) are ordered within each paradigm and modeling scenario when configurations are sorted by BA, using the recorded metrics from Sections 5.2 and 5.3. The rankings in Table 5.8 are obtained mechanically by applying the ranking procedure described in Section 5.10 to the reported BA values (and using other core metrics as tie-breakers where applicable). No probabilistic or inferential claims are attached to these rankings.

Paradigm	Modeling Scenario	Ranking within scenario (by BA)
MI	Classical ML (LDA/LR)	Spectrogram ≳ PSD > PLV > ApEn
	DL (CNN/LSTM, raw EEG)	Raw–CNN > Raw–LSTM, according to the BA values reported in Table 5.4.
ERP	Classical ML (LDA/LR)	Spectrogram > PLV ≳ PSD > ApEn
	DL (CNN/LSTM, raw EEG)	Raw–CNN > Raw–LSTM, according to the BA values reported in Table 5.5.
SSVEP	Classical ML (LDA/LR)	Spectrogram > PSD > PLV > ApEn
	DL (CNN/LSTM, raw EEG)	Raw–CNN > Raw–LSTM, according to the BA values reported in Table 5.6.

Table 5.8: Qualitative rankings obtained by ordering configurations according to Balanced Accuracy BA within each paradigm and modeling scenario, based solely on the metrics reported in previous tables. Symbols (>, ≳) denote ordering relations induced by the recorded BA values and do not represent statistical significance.

5.5 Computational Footprint and Practical Trade-offs

This section reports qualitative information about the computational characteristics of the evaluated pipelines. The focus is on separating feature-extraction cost from model-inference cost under the implementations described in Chapter 4. Exact runtimes and parameter counts were not logged in a uniform manner across all configurations; therefore, only relative and protocol-based descriptions are provided. No additional metrics beyond those already reported in previous sections are introduced here.

5.5.1 ML Baselines

For the classical models LDA and LR applied to ApEn, PLV, PSD , and Spectrogram features:

- Training was performed using standard convex optimization routines on the pre-computed feature matrices for each dataset–feature combination. Under the dataset sizes considered in this thesis, training completed within short wall-clock times on standard Central Processing Unit (CPU) hardware, according to the experimental logs.

- Inference consists of evaluating linear decision functions (and softmax outputs for LR) on feature vectors. This operation has negligible per-trial latency relative to typical EEG sampling rates.

- For these pipelines, the main additional cost arises from feature computation: PSD and Spectrogram features require spectral/time–frequency estimation per trial, and PLV features require computing pairwise relationships across channels as implemented in Chapter 4.

No changes to the reported performance metrics are associated with these descriptions. They provide contextual information about how the classical baselines were obtained.

5.5.2 DL Models

For the DL backbones (CNN and LSTM) applied to Raw:

- The architectures used in this thesis are compact CNN and LSTM models with a moderate number of layers and parameters, as specified in Chapter 4. All configurations were trainable on a single Graphics Processing Unit (GPU) or CPU for the dataset sizes used.

- For convolutional models, inference consists of a sequence of convolution, normalization, pooling, and fully connected layers applied to the chosen input representation (e.g., Raw time series or time–frequency maps). For the evaluated input lengths and channel counts, per-trial inference time remained within ranges suitable for offline and potential online use under the experimental conditions.

- For LSTM-based models, inference proceeds sequentially over time steps for each trial. Under the epoch durations used in these experiments, this was also feasible within typical computational constraints reported for the environment in which the experiments were executed.

These descriptions are intended to document the qualitative computational profile of the evaluated deep configurations. They do not modify or re-interpret the quantitative performance results already reported in Sections 5.2 and 5.3.

5.6 Summary

Chapter 5 reports the experimental outcomes for all combinations of:

- three benchmark EEG paradigms and datasets (MI, ERP, SSVEP),
- four feature or input representations (PSD , PLV, ApEn, Spectrogram)and Raw EEG.
- and four model types (LDA, LR, CNN, LSTM),

evaluated under a subject-aware, leakage-free protocol as defined in Chapter 4.

For each configuration, results are presented using a fixed set of core metrics: BA, Macro-F1, Cohen's κ, AUROC_macro, and AUPRC_macro.

The present chapter has compiled all recorded configuration-level results under a unified evaluation protocol, without adding interpretation or methodological modification. These empirical outcomes define the quantitative basis upon which the following chapter builds. In the next chapter, these results are examined comparatively and conceptually, with a structured discussion of their implications for feature representations, model choices, and the broader design of EEG-based decoding frameworks.

Chapter 6

Discussion

This chapter interprets the empirical results reported in Chapter 5 in the light of the RRP, the theoretical background established in Chapter 2, and recent advances in EEG-based decoding and semantic feature extraction. The focus is on what the observed configuration-level outcomes imply about how information is organized in EEG signals, how different representations expose or obscure that information, and how these choices interact with model class and evaluation protocol.

6.1 Revisiting the RRP

The RRP, introduced in Section 1.3, states that, for a fixed task and computational budget, feature spaces that preserve more of the task-relevant spatio-temporal–spectral structure of EEG should support more reliable and more generalizable decoding than heavily compressed, low-capacity summaries, provided the decoding model has suitable inductive biases.Figure 6.1.

Across all datasets, the strongest configurations consistently involve representations that either retain the raw multichannel waveform or encode structured spectral or time–frequency information, combined with models that can exploit that structure CNNs. In contrast, aggressively compressed complexity-only features ApEn and narrowly defined connectivity-only descriptors PLV do not approach the performance of richer alternatives, even when paired with deep models. This pattern supports the central premise of the RRP: preserving structure is a necessary condition for high-capacity decoders to access task-relevant information in EEG.

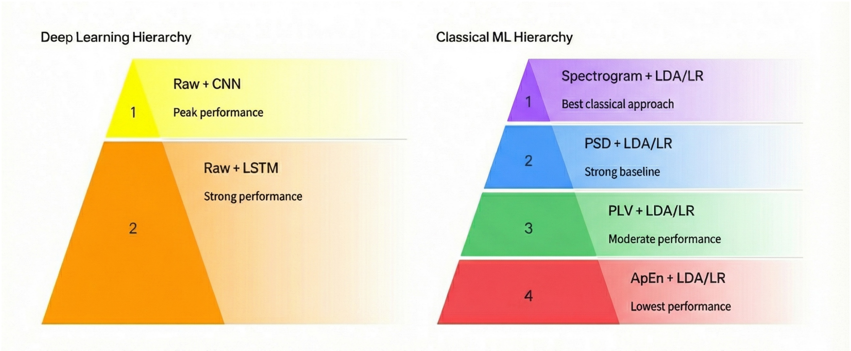

Figure 6.1: Rankings obtained by ordering configurations according to BA.

6.2 Semantic Capacity of Different Representations

6.2.1 Raw

The Raw+CNN configurations in all three paradigms provide some of the best BA, macro-F1, κ, AUROC_macro, and AUPRC_macro scores reported in Chapter 5. These outcomes are consistent with the view from Chapter 2 that semantically relevant information is embedded in distributed, multiscale activity patterns, and that preserving the full spatio temporal signal allows a model to discover discriminative structure without being constrained by strong hand-crafted compression.

6.2.2 Connectivity and Complexity-Only Features

PLV and ApEn were included as deliberately constrained representations: PLV emphasizes phase relationships; ApEn emphasizes scalar complexity. The results in Chapter 5 show that: classical models on PLV/ApEn, and deep models operating solely on these features, consistently underperform relative to PSD, and Spectrogram pipelines across all paradigms.

This outcome does not negate the potential utility of connectivity or complexity measures as part of a richer representational stack. Rather, it indicates that when used in isolation and aggressively summarized, they provide insufficient semantic capacity for robust decoding across heterogeneous subjects and tasks. In the language of Section 1.3, these feature spaces occupy the bottom rungs of the richness ladder: they are too restrictive to serve as universal substrates for task-level meaning and intention.

6.3 Model Class as a Moderator of Richness

6.3.1 ML Models

LDA and LR serve in this thesis as conservative, transparent baselines. Their performance on PSD and Spectrogram features, particularly for ERP and SSVEP, confirms that a substantial portion of task-relevant structure becomes linearly accessible once appropriate spectral or time–frequency representations are constructed. However, their ceiling is clearly visible in Chapter 5: even under the best feature choices, they do not match the top-performing deep configurations that operate on richer inputs.

In the context of RRP, linear baselines play two roles: (i) they demonstrate how far one can go with low-variance decision boundaries on engineered features; (ii) they provide a reference line against which any claimed advantage of deep models or new representations must be measured.

6.3.2 DL Models

The deep backbones considered here were intentionally modest in size to keep the comparison focused on inductive bias rather than scale. Across paradigms, CNNs consistently outperform LSTMs on the same input.This suggests that convolutional architectures with local, hierarchical filters align more naturally with the spatial and temporal organization of ERP, MI, and SSVEP signals than generic recurrent sequence models.

6.4 Representation Model Matching and the Limits of Naive Richness

A central observation arising from the results in Chapter 5 is that “richness” cannot be defined solely by the apparent dimensionality or layout of a feature. A matrix-valued or tensor-valued representation is not automatically more informative than a vector, and connectivity maps are not inherently superior to band-power profiles independent of the model that consumes them.

In practice, the effective semantic capacity of a representation is determined by the *pair* (representation, model). A feature map is only rich to the extent that its structure can be exploited by the inductive biases and optimization dynamics of the model. Convolutional networks are adapted to local spatio temporal patterns; linear models are adapted to globally separable projections; recurrent models emphasize sequential dependencies. A representation that encodes information in forms misaligned with these biases may have high apparent complexity but low functional utility.

The configuration-level outcomes reported in Chapter 5 illustrate this interaction: structurally elaborate features such as PLV or ApEn do not automatically outperform simpler spectral or time–frequency descriptors, and increasing model capacity alone does not guarantee improved performance on such inputs. Conversely, Raw representations yield strong results precisely when paired with architectures (e.g., CNNs) that can absorb and organize their spatio–temporal and spectral structure. Because modern deep architectures remain effectively black boxes, this matching is only partially characterizable a priori; in most cases it must be established empirically rather than assumed as a universal design rule.

Therefore, even though many of the present findings are consistent with the intuition that structure-preserving representations tend to be more informative, they do not justify a universal prescription such as "matrix-valued features are always better than vectors." The mapping from raw EEG to an effective feature space remains contingent on model class, task demands, recording conditions, and evaluation protocol.

This limitation motivates the move beyond narrowly engineered, paradigm-specific features toward representations that are (i) constructed or learned to capture task-relevant meaning across heterogeneous conditions and (ii) less sensitive to ad hoc representation–model pairings. In this context, task-independent and semantically oriented frameworks—such as topographic sequence encodings [13] and the universal semantic feature extraction framework of Ahmadi and Mesin [14]—provide promising targets: they aim to define embedding spaces in which diverse EEG tasks can be decoded using a shared representational backbone, rather than relying on fragile, case-specific feature–model matches.

6.5 Cross-Paradigm Behaviour and Generalization

The three paradigms sampled in this thesis (MI, ERP, and SSVEP) probe different neural mechanisms and signal structures: oscillatory modulation over sensorimotor cortex, ERP and steady-state responses to periodic visual stimulation. Despite these differences, several stable patterns emerge:

- In all paradigms, Raw+CNN and appropriately structured spectral or time–frequency inputs yield the highest configuration-level metrics.
- ML models on PSD/Spectrogram form meaningful, but clearly suboptimal, reference points.
- PLV and ApEn-only configurations consistently rank at the bottom across BA, Macro-F1, κ, AUROC_macro, and AUPRC_macro.

This consistency across paradigms indicates that the observed hierarchy of representations is not specific to a single protocol or label space. Instead, it reflects a more general relationship between preserved structure and semantic capacity under subject-aware evaluation: configurations that expose richer spatio–temporal–spectral organization of EEG tend to yield more stable decoding across heterogeneous conditions.

6.6 Positioning with Respect to Prior Work

6.6.1 High-Performing, Dataset-Specific Pipelines

A substantial body of recent work has reported very high within-dataset performance for specific paradigms using sophisticated ensembles and deep models. For example, Ahmadi and Mesin introduced the COWSE for MI EEG in [8], and related stacked adaptive ensemble approaches for MI classification across multiple datasets in [9], achieving accuracies that approach or exceed conventional benchmarks. Similarly, hierarchical deep models for coma outcome prediction have reported near-perfect discrimination on specific cohorts [10], and advanced architectures for secure and robust brain-to-brain communication have demonstrated strong robustness and confidentiality guarantees on targeted SSVEP and BCI settings [11, 12].

These studies collectively show that, given a fixed dataset and careful tuning, high-capacity models and ensembles can fit the available structure extremely well. However, as also emphasized in Chapter 2, direct transfer of such pipelines to different datasets, recording setups, or subject populations often yields substantially degraded performance. The results in Chapter 5, obtained under a unified, subject-aware and leakage-free protocol across multiple paradigms, complement this literature by: (i) placing classical and deep approaches in a common evaluation frame, and (ii) attributing observed gains or failures not just to architectures, but to the representational choices that either preserve or discard task-relevant structure.

6.7 Practical Implications for BCI and EEG Decoding

The combined evidence from Chapter 5 and the literature suggests several practical guidelines for designing EEG decoding pipelines Figure 6.2

- Rich, structure-preserving inputs (Raw, well-designed time–frequency maps, or embeddings inspired by universal/semantic frameworks) should be considered first-class options when computational resources allow.
- ML models remain useful as low-cost, interpretable baselines, particularly on engineered spectral features, but they do not fully exploit the semantic capacity of richer representations.
- Deep architectures should be paired with representations that expose meaningful structure; applying complex models to aggressively compressed or narrowly defined features offers limited benefit.
- Unified, subject-aware, and leakage-free evaluation protocols are essential for making credible claims about generalization, especially in light of previously reported near-ceiling accuracies on single datasets.

These implications resonate with recent task-independent approaches such as topomap sequences and universal semantic feature extraction [13, 14], which operationalize representation richness in more explicit ways.

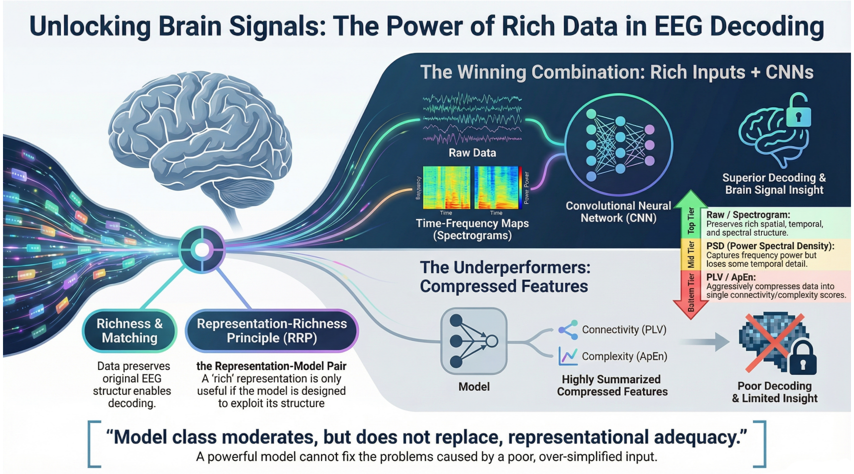
Unlocking Brain Signals: The Power of Rich Data in EEG Decoding
The Winning Combination: Rich Inputs + CNNs
Raw Data
Time
Time
Time-Frequency Maps (Spectrograms)
Convolutional Neural Network (CNN)
Superior Decoding & Brain Signal Insight
Top Tier
Raw / Spectrogram: Preserves rich spatial, temporal, and spectral structure.
Mid Tier
PSD (Power Spectral Density): Captures frequency power but loses some temporal detail.
Baltem Tier
PLV / ApEn: Aggressively compresses data into single connectivity/complexity scores.
The Underperformers: Compressed Features
Richness & Matching
Data preserves original EEG structur enables decoding.
Representation-Richness Principle (RRP)
the Representation-Model Pair
A 'rich' representation is only useful if the model is designed to exploit its structure
Model
Connectivity (PLV)
Complexity (ApEn)
Highly Summarized Compressed Features
Poor Decoding & Limited Insight
"Model class moderates, but does not replace, representational adequacy."
A powerful model cannot fix the problems caused by a poor, over-simplified input.

Figure 6.2: Conceptual framework contrasting Rich Inputs against Compressed.

6.8 Methodological Considerations and Limitations

Several methodological aspects constrain the scope of the conclusions:

- The study focuses on three paradigms and a specific set of public datasets. While they are widely used benchmarks, they do not cover all cognitive, affective, or clinical EEG use-cases.

- Only four representation with Raw EEG signals regimes and four model classes are evaluated. More complex architectures (e.g., transformers, graph neural networks, contrastive representation learning) and hybrid feature constructions fall outside the present experimental design.

- Results are reported at configuration level; complete per-subject distributions were not available for all settings. Consequently, the chapter relies on descriptive comparisons rather than formal statistical inference.

- Hyperparameters and architectures are intentionally kept fixed across datasets to preserve comparability. This choice emphasizes systematic trends but may understate the best achievable performance of certain models under paradigm-specific tuning.

These constraints are consistent with the thesis goal: to test the Representation–Richness Principle under controlled and realistic, but not exhaustive, conditions.

6.9 Future Directions

While the RRP confirms that preserving spatiotemporal structure is crucial, 'structure' alone does not guarantee that the extracted features map one-to-one with the user's cognitive intent. Current rich representations (like Raw EEG or Spectrograms) are still signal-centric rather than meaning-centric. They describe how the signal looks, not necessarily what it means.

This limitation necessitates a paradigm shift towards Semantic Feature Extraction. Unlike traditional feature engineering that targets signal properties (amplitude, phase, entropy), semantic features aim to learn an embedding space where the geometric relationships between vectors reflect the conceptual relationships between the underlying tasks. Moving from purely structural richness to semantic richness represents the logical next step to overcome the cross-paradigm generalization challenges observed in this study."

This study deliberately restricts model complexity, aggressive tuning, and architecture specialization to preserve a level playing field, directing attention to how different EEG representations themselves shape performance. While this design choice may yield absolute results below highly optimized state-of-the-art systems for specific paradigm–feature combinations, it strengthens the interpretability and robustness of the *relative* comparisons across feature families and between ML and DL approaches, reducing the risk of overfitting through exhaustive hyperparameter searches.

Building on this foundation, future work can integrate richer EEG-specific architectural priors (such as graph-based models for connectivity representations), explore joint multi-view frameworks that fuse raw and engineered features under shared calibration constraints, and systematically assess domain adaptation strategies for cross-session, cross-subject, and cross-device transfer. Extending analyses of probabilistic calibration and threshold-free separability will further support rigorous evaluation of reliability and trustworthiness, a prerequisite for deploying these models in real-world BCI applications.

Recent contributions by Ahmadi, Mesin, and collaborators explicitly move toward task-independent, structure-preserving EEG representations. The topomap-sequence framework for decoding visual imagination and perception treats EEG as a sequence of scalp maps and applies spatio–temporal models to this image-like representation [13], aligning closely with the richness-oriented stance of this thesis. Even more directly, the “Universal Semantic Feature Extraction from EEG Signals: A Task-Independent Framework” [14] proposes a unified embedding that aims to capture semantic structure across heterogeneous datasets and paradigms. Conceptually, such universal embeddings occupy the upper end of the richness ladder: they are designed to preserve distributed, relational, and potentially conceptual information in a shared feature space, rather than being tied to a single protocol.

The present thesis is complementary to these advances. Instead of proposing a single new universal representation, it systematically benchmarks a spectrum of widely used feature regimes and model classes under controlled conditions. The results support the same general intuition underlying these newer frameworks: representations that more faithfully preserve the structure of EEG signals are better candidates for capturing task-general and semantically meaningful information than heavily compressed traditional descriptors.

Building on the present findings and recent work by Ahmadi, Mesin, and others [8, 9, 10, 11, 12, 13, 14], several concrete extensions suggest themselves:

- **Universal embeddings.** Develop and evaluate task-independent EEG embeddings, trained jointly on heterogeneous paradigms, and compare them directly against the representation ladder used in this thesis, including systematic tests of their semantic structure and calibration behaviour.
- **Multi-view and hybrid features.** Combine raw, time–frequency, connectivity (e.g., graph-based) and semantic embeddings in unified, multi-view architectures to test whether complementary structure further improves robustness, calibration, and threshold-free separability.
- **Cross-dataset and cross-lab transfer.** Extend the evaluation to additional datasets, recording conditions, and hardware, with stringent domain shift (cross-session, cross-subject, and cross-device) to test whether rich representations indeed capture stable, transferable organization and to benchmark domain adaptation strategies.
- **Calibration, reliability, and security.** Systematically link representation choices to calibration quality, uncertainty estimates, and vulnerability or robustness in adversarial, privacy-sensitive, and brain-to-brain communication scenarios, extending analyses of probabilistic calibration and threshold-free separability toward reliability criteria suitable for real-world BCI deployment.

Taken together, these directions aim toward EEG decoding systems whose behaviour is both empirically grounded and conceptually aligned with distributed neural coding of meaning and intention.

Chapter 7

Conclusion

This thesis systematically investigated how EEG representations and model classes jointly determine the reliability and semantic plausibility of decoding under realistic, subject-aware constraints. The aim was to provide a controlled empirical test of the practical RRP: whether representations that preserve more temporal, spectral, and spatial structure in the signal support more meaningful task-level decoding than heavily compressed descriptors.

The empirical findings confirm that configurations relying on rich, structure-preserving representations (such as Raw EEG or Spectrograms) consistently outperform those limited to compressed scalars. Nevertheless, traditional features, often reliant on manual extraction or superficial data properties, inherently possess insufficient representational power to effectively capture the latent hidden patterns required for complex tasks. In contrast, DL models have offered a significant improvement by learning hierarchical, automatic features that yield better performance than their traditional counterparts.

Although richer features consistently yielded better results in our experiments, the persistent performance variability across subjects and paradigms highlights a remaining gap. We observed that richness is a necessary condition—providing the model with enough information—but it is not sufficient on its own. The feature space and the model must be synergistically aligned not just to the signal's topology, but to the latent cognitive constructs.

Consequently, this thesis points to Semantic Feature Extraction as the essential frontier for future BCI research. By building upon the structural preservation advocated by the RRP, future systems must move beyond simply decoding signal statistics to encoding the conceptual meaning of the data. This transition—from capturing signal complexity to capturing semantic intent—promises to bridge the generalization gap identified in this work, paving the way for intelligent systems that truly 'understand' neural activity rather than merely classifying it.

Bibliography

[1] Marc R. Nuwer. “10-10 electrode system for EEG recording”. In: *Clinical Neurophysiology* 129.5 (2018), p. 1103. ISSN: 1388-2457. DOI: https://doi.org/10.1016/j.clinph.2018.01.065. URL: https://www.sciencedirect.com/science/article/pii/S1388245718300907 (cit. on p. 3).

[2] Tom M. Mitchell et al. “Predicting human brain activity associated with the meanings of nouns”. In: *Science* 320.5880 (2008), pp. 1191–1195. DOI: 10.1126/science.1152876 (cit. on p. 9).

[3] Alexander G. Huth et al. “Natural speech reveals the semantic maps that tile human cerebral cortex”. In: *Nature* 532 (2016), pp. 453–458. DOI: 10.1038/nature17637 (cit. on p. 9).

[4] Alexandre Barachant, Sylvain Bonnet, Marco Congedo, and Christian Jutten. “Multiclass Brain–Computer Interface Classification by Riemannian Geometry”. In: *IEEE Trans. Biomed. Eng.* 59.4 (2012), pp. 920–928. DOI: 10.1109/TBME.2011.2172210 (cit. on pp. 10, 11).

[5] Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. “Deep Learning With Convolutional Neural Networks for EEG Decoding and Visualization”. In: *Human Brain Mapping* 38.11 (2017), pp. 5391–5420. DOI: 10.1002/hbm.23730 (cit. on p. 12).

[6] Vernon J. Lawhern et al. “EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces”. In: *J. Neural Eng.* 15.5 (2018), p. 056013. DOI: 10.1088/1741-2552/aace8c (cit. on p. 12).

[7] Vinay Jayaram and Alexandre Barachant. “MOABB: trustworthy algorithm benchmarking for BCI”. In: *Proc. 7th Graz Brain-Computer Interface Conference*. 2017 (cit. on p. 12).

[8] H. Ahmadi and L. Mesin. “Enhancing Motor Imagery Electroencephalography Classification with a Correlation-Optimized Weighted Stacking Ensemble Model”. In: *Electronics* 13 (2024), p. 1033. DOI: 10.3390/electronics13061033 (cit. on pp. 13, 66, 69).

[9] H. Ahmadi and L. Mesin. “Enhancing MI EEG Signal Classification With a Novel Weighted and Stacked Adaptive Integrated Ensemble Model: A Multi-Dataset Approach”. In: *IEEE Access* 12 (2024), pp. 103626–103646. DOI: 10.1109/ACCESS.2024.3434654 (cit. on pp. 13, 66, 69).

[10] H. Ahmadi, P. Costa, and L. Mesin. *A Novel Hierarchical Binary Classification for Coma Outcome Prediction Using EEG, CNN, and Traditional ML Approaches*. TechRxiv. 2024. DOI: 10.36227/techrxiv.173220750.05940724 (cit. on pp. 13, 67, 69).

[11] H. Ahmadi, A. Kuhestani, and L. Mesin. “Adversarial Neural Network Training for Secure and Robust Brain-to-Brain Communication”. In: *IEEE Access* 12 (2024), pp. 39450–39469. DOI: 10.1109/ACCESS.2024.3376657 (cit. on pp. 13, 67, 69).

[12] H. Ahmadi, A. Kuhestani, M. Keshavarzi, and L. Mesin. “Securing Brain-to-Brain Communication Channels Using Adversarial Training on SSVEP EEG”. In: *IEEE Access* 13 (2025), pp. 14358–14378. DOI: 10.1109/ACCESS.2025.3528770 (cit. on pp. 13, 67, 69).

[13] H. Ahmadi and L. Mesin. *Decoding Visual Imagination and Perception from EEG via Topomap Sequences*. TechRxiv. 2025. DOI: 10.36227/techrxiv.174672922.22051031 (cit. on pp. 14, 66, 67, 69).

[14] H. Ahmadi and L. Mesin. “Universal semantic feature extraction from EEG signals: a task-independent framework”. In: *J. Neural Eng.* 22.3 (2025). DOI: 10.1088/1741-2552/add08f (cit. on pp. 15, 66, 67, 69).