

Subject-Independent Semantic Decoding from EEG: Comparing Interpretable Feature-Based, Machine-Learning, and Deep-Learning Approaches

Vittorio Fonte (s265954)
Master Degree in Biomedical Engineering (2025/2026)
Politecnico di Torino

Supervisor: Prof. Luca Mesin
Co-supervisor: PhD Hossein Ahmadi

1 Introduction

Semantic decoding aims to identify, directly from brain activity and without any motor or behavioural response, which concept a person is perceiving or imagining. Electroencephalography (EEG) is the most widely adopted recording modality in Brain-Computer Interface (BCI), owing to its non-invasiveness, low cost, portability, and high temporal resolution, yet its use for semantic decoding remains an emerging line of research, hindered by low signal-to-noise ratio, temporal non-stationarity, muscular and ocular artefacts, and limited spatial resolution. Two limitations pervade the field. The first is poor cross-subject generalisation: within-subject and pooled evaluation protocols reward subject-specific idiosyncrasies rather than stimulus-driven representational structure, yielding optimistic estimates and decoders that are effectively subject-specific. The second is limited interpretability: the field is increasingly dominated by opaque deep-learning models that resist direct neuroscientific interpretation. This work is situated in that gap. Using the multimodal semantic EEG dataset of Wilson et al. (the first to capture semantic representation across several sensory modalities and for both perception and imagination within a single sample), it applies a strict Leave-One-Subject-Out (LOSO) protocol to three classification tasks: perception vs imagination, sensory modality (auditory, orthographic, pictorial), and semantic concept (flower, penguin, guitar). For the last two tasks, this is the first evaluation under a subject-independent regime. The central contribution is a novel interpretable pipeline that couples an anatomical regional decomposition with region-wise Common Spatial Patterns (CSP) and a parameter-free decision rule. The benchmark is performed against two comparators: a machine-learning approach based on Riemannian geometry, and a deep-learning approach based on topographic decoding, thus establishing subject independency while complementing established approaches. Three research questions guide the work:

- **RQ1:** does the variance of the class-specific CSP component constitute a discriminative feature under a subject-independent regime, and does the proposed decorrelation measure add value beyond uniform regional energy aggregation?
- **RQ2:** can Riemannian covariance features sustain performance across the three tasks under LOSO?
- **RQ3:** does topographic decoding, already effective on this dataset for the pictorial perception–imagination distinction, generalise to the remaining tasks?

2 Materials and Methods

Dataset. The Wilson et al. dataset (OpenNeuro, BIDS format) comprises 12 participants across 15 sessions (nine with one session, three with two), each lasting about two hours, recorded with a 124-channel ANT Neuro scalp montage (1024 Hz, 24-bit, 10–5 system, CPz reference). The three concrete

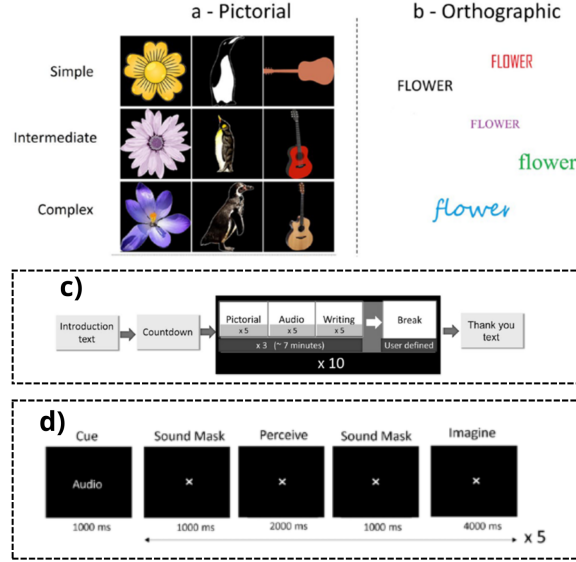


Figure 1: a) and b) show the intra-class variance of the same stimulus-class. c) Examples of visual stimuli used in the experiment. c) Overview of the session structure. d) Example of the experimental protocol

concepts are chosen to be maximally distinct in meaning, while matched on low-level attributes (identical syllable count, deliberately uncommon objects). Six paradigms arise from crossing two cognitive tasks (perception, imagination) with three sensory modalities (pictorial, orthographic, auditory); the pronounced within-category variability (multiple exemplars, complexity levels, colors, fonts, voices) forces the decoder to rely on semantic content (Figure 1). Owing to fatigue the number of trials per condition varies (64–150); analysis windows last 4 s for imagination and 2 s or 3 s, for auditory and visual perception respectively.

Preprocessing. The preprocessing released by the authors is adopted unchanged (bad-channel interpolation, common-average referencing, notch filtering at 50/100/150 Hz and a 2 Hz high-pass, artefact removal by FastICA), followed by per-channel z-scoring.

Pipeline 1: Interpretable Handcrafted Approach. The 124 electrodes, excluding the four mid-line ones, are partitioned into 10 anatomical regions of interest (12 electrodes each). Within each region, region-wise CSP filters are applied (One-vs-Rest for the three-class tasks), from which two interpretable scalar descriptors are extracted: a regional energy (variance of the first CSP component) and an “attention” or decorrelation measure (the complement of the absolute cosine similarity between the two CSP components). Classification proceeds through parameter-free decision rules (energy only, attention only, or their multiplicative combination), the CSP filters being the only quantities learned from data. The pipeline operates on the full preprocessed band.

Pipeline 2: Machine Learning Comparator (Riemannian Geometry). Each trial is characterised by its spatial covariance matrix, estimated with Ledoit–Wolf shrinkage separately in five physiological bands (δ , θ , α , β , γ). The covariances, which lie on the manifold of SPD matrices, are projected onto the tangent space at the Fréchet mean of the training set (affine-invariant metric); the five tangent vectors are concatenated ($5 \times 7750 = 38,750$ dimensions per trial) and fed to two supervised classifiers, a linear SVM and a random forest.

Pipeline 3: Deep Learning Comparator (Topographic Decoding). Adapting the approach of Ahmadi and Mesin, a representative epoch is built for each subject and class by averaging that subject’s trials sample-by-sample (in analogy with event-related potentials). The epoch is converted

into a sequence of scalp-voltage maps (topomaps) at 10 ms intervals, rendered as 32×32 RGB images; each frame is a classification instance (frame-as-sample) fed to a compact convolutional network.

Evaluation. All pipelines are evaluated under LOSO with per-session balanced accuracy (15 folds). The statistical analysis is non-parametric, namely a one-sided Wilcoxon signed-rank test against chance with Benjamini–Hochberg (BH) correction. For inter-pipeline comparison, a Friedman omnibus test precedes pairwise Wilcoxon post-hoc tests, again with a BH correction.

3 Results

Imagination vs Perception. Pipeline 1 never exceeds chance, with a balanced accuracy close to 0.52. The two data-driven approaches decode the distinction well above chance: Pipeline 2 (random forest) is the strongest, reaching 0.94 balanced accuracy for auditory stimuli and 0.79 for orthographic and pictorial. Conversely, pipeline 3 (CNN) is robustly above chance and more uniform (balanced accuracy up to 0.70). The Pipeline 2 reach the highest scores in every modality, with particular statistical significance for auditory and visual orographic.

Sensory Modality. Pipeline 1 remains at chance in both conditions. Pipeline 2 discriminates genuinely only in perception, with the highest score for the auditory class, whereas in imagination it is barely above chance. Pipeline 3 exceeds chance in both (imagination 0.47, perception 0.54) and recovers the three modalities more evenly. Overall, pipeline 2 and 3 exceed pipeline 1 in perception, whereas, pipeline 3 scores better in imagination.

Semantic Concept. This is the task on which the pipelines diverge the most. Neither pipeline 1 nor pipeline 2 exceed chance in any of the six task–modality conditions. On the contrary, Pipeline 3 exceeds chance in all six (with balanced accuracy ranging from 0.46 to 0.52): it is the only method to recover semantic information under a subject-independent regime.

4 Conclusion

In this work, three different pipelines for EEG decoding have been evaluated on three problems of increasing semantic depth, under a strict Leave-One-Subject-Out protocol; performance was governed jointly by the task and by each pipeline’s representation, and no single method dominated throughout.

RQ1: Handcrafted Regional CSP. Under the parameter-free rules adopted, the class-specific CSP variance did not transfer across subjects: Pipeline 1 never exceeded chance, clearing it only once and marginally (auditory perception–imagination), and the proposed decorrelation measure (the attention) does not improve over uniform energy aggregation. The result delimits what a minimally configured interpretable CSP can achieve, as the chosen configuration includes all the frequency components of the signal without any specific bands filtering prior to applying the CSP, possibly at the expense of reduced performance.

RQ2: Riemannian Covariance Features. The pipeline 2 proved to be a viable approach for decoding the perception–imagination distinction in all three modalities. Although, performance limitations were observed in harder problems, specifically when recognizing semantic categories. Such behavior may originate from the high ratio of the number of features with respect to trials. Additionally, it may be hypothesized that in high level cognitive tasks, not all the EEG bands are equally important, thus, acting as noise sources.

RQ3: Topographic Decoding. The pipeline 3 is the only method to generalize to harder classification tasks. When using averaged trials per subject, the signal noise ratio increases only if there is a common pattern, which facilitate the classification task. Despite the limited data available, the result suggests that such an underlying pattern may actually exists, allowing for further experimentation in semantic tasks.