Universal semantic feature extraction from EEG signals: a task-independent framework

PAPER • **OPEN ACCESS**

Universal semantic feature extraction from EEG signals: a task-independent framework

Journal of Neural Engineering

PAPER

Universal semantic feature extraction from EEG signals: a task-independent framework

Hossein Ahmadi* and **Luca Mesin***

Mathematical Biology and Physiology, Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy
* Authors to whom any correspondence should be addressed.

E-mail: hossein.ahmadi@polito.it and luca.mesin@polito.it

Keywords: semantic feature extraction, task-independent features, neural representations, EEG decoding, Transformer

OPEN ACCESS

RECEIVED
7 January 2025

REVISED
31 March 2025

ACCEPTED FOR PUBLICATION
24 April 2025

PUBLISHED
6 May 2025

Abstract

Objective. Extracting universal, task-independent semantic features from electroencephalography (EEG) signals remains an open challenge. Traditional approaches are often task-specific, limiting their generalization across different EEG paradigms. This study aims to develop a robust, unsupervised framework for learning high-level, task-independent neural representations. *Approach.* We propose a novel framework integrating convolutional neural networks, AutoEncoders, and Transformers to extract both low-level spatiotemporal patterns and high-level semantic features from EEG signals. The model is trained in an unsupervised manner to ensure adaptability across diverse EEG paradigms, including motor imagery (MI), steady-state visually evoked potentials (SSVEPs), and event-related potentials (ERPs, specifically P300). Extensive analyses, including clustering, correlation, and ablation studies, are conducted to validate the quality and interpretability of the extracted features. *Main results.* Our method achieves state-of-the-art performance, with average classification accuracies of 83.50% and 84.84% on MI datasets (BCICIV_2a and BCICIV_2b), 98.41% and 99.66% on SSVEP datasets (Lee2019-SSVEP and Nakanishi2015), and an average AUC of 91.80% across eight ERP datasets. t-distributed stochastic neighbor embedding and clustering analyses reveal that the extracted features exhibit enhanced separability and structure compared to raw EEG data. Correlation studies confirm the framework's ability to balance universal and subject-specific features, while ablation results highlight the near-optimality of the selected model configuration. *Significance.* This work establishes a universal framework for task-independent semantic feature extraction from EEG signals, bridging the gap between conventional feature engineering and modern deep learning methods. By providing robust, generalizable representations across diverse EEG paradigms, this approach lays the foundation for advanced brain–computer interface applications, cross-task EEG analysis, and future developments in semantic EEG processing.

1. Introduction

Electroencephalography (EEG) has long provided crucial insights into brain activity, relying on features such as spectral power, band-specific energy, and spatial connectivity to understand cognitive processes and behaviors [1, 2]. While these traditional features have significantly advanced the field, they often capture only low-level, task-specific characteristics, overlooking the deeper semantic layers of meaning, intention, and context encoded in neural signals.

The extraction of universal, task-independent semantic features from EEG data remains an underexplored frontier. Existing approaches that touch upon semantic interpretations tend to be narrowly tailored to specific tasks or depend on supervised methods, limiting their generality across different contexts and subjects. Developing a universal semantic feature would not only enrich our theoretical understanding of how the brain encodes meaning but also open up practical avenues for brain–computer interfaces (BCIs), brain-to-brain

communication (B2B-C) systems, neurodiagnostics, education, and industrial automation.

Achieving this goal is challenging due to the complexity and variability inherent in neural representations, as well as the influence of contextual factors. Yet, enabling a broader, more flexible analysis of EEG signals is necessary. To address this challenge, our research focuses on three main objectives:

1. Defining what constitutes a semantic feature in EEG data,
2. outlining criteria and expectations for a universal semantic feature, such as task independence, robustness to inter-subject variability, and meaningfulness,
3. and assessing the feasibility of achieving these aims using state-of-the-art, unsupervised, and self-supervised deep learning (DL) techniques.

While our endeavor is ambitious, our approach is incremental. We propose and validate a framework to progressively capture semantic content, bridging the gap between the need for semantic interpretation and the limitations of current EEG feature sets. By doing so, we aim to establish a universal semantic representation independent of specific tasks or conditions, thus facilitating a wide range of analyses, including supervised classification, unsupervised clustering, cross-subject comparisons, and within-subject investigations over time.

Ultimately, this work lays the groundwork for transforming EEG analysis from a task-centric paradigm to one capable of universal semantic interpretation. Such progress is essential for advancing neuroscience research, fostering interdisciplinary collaborations, and enhancing the capabilities of neurotechnology and neuroinformatics applications.

The organization of this paper is as follows: section 2 discusses related work, followed by the proposed methodology in section 3. Section 4 provides the mathematical formulation, and section 5 presents the results. Section 6 discusses the findings in detail, and finally, section 7 concludes the study with key insights and future directions.

2. Related works

EEG signal processing has progressed substantially, yielding successful applications in diverse tasks. Ensemble learning frameworks for motor imagery (MI) classification, such as correlation-optimized weighted stacking ensembles and weighted and stacked adaptive integrated ensemble classifiers, have improved accuracy and robustness across multiple datasets [3, 4]. Advanced DL models have been developed in clinical settings to predict neurologic outcomes in comatose patients. For instance, a multiscale deep neural network combining convolutional neural networks (CNNs) and long short-term memory (LSTM) networks has been utilized to analyze EEG data and demographic information, achieving an area under the receiver operating characteristic curve (AUC-ROC) of up to 0.91 within 66 h post-cardiac arrest. This approach underscores the efficiency and practical relevance of integrating CNNs with temporal dynamics for accurate coma outcome prediction [5].

Efforts to improve the security and robustness of EEG-based systems have introduced frameworks that enhance the resilience of B2B-C systems against adversarial attacks through adversarial neural network training. These approaches have optimized architectures, such as CNN-temporal convolutional networks, achieving significant improvements in accuracy and security metrics [6, 7].

Efforts toward semantic feature extraction from EEG signals have emerged in various domains. Studies examining potency, valence, and arousal via high-density EEG and source localization have revealed distinct spatiotemporal patterns for abstract and concrete words, illustrating EEG's ability to represent subtle semantic distinctions [8]. Similarly, differential entropy and peak-magnitude root mean square ratio, combined with variational mode decomposition, have improved epileptic seizure classification through semantic feature extraction [9]. Event-related potential (ERP) investigations (e.g. N400 amplitude) further highlight that semantic relevance, rather than feature type or category, influences neural responses during concept retrieval [10].

Recent approaches integrate EEG with visual data for semantic-driven image reconstruction. An EEG-visual-residual-network and diffusion models have jointly achieved high-resolution image generation [11], while dual conditional AutoEncoders fusing multimodal semantic features with UNet-based image generation have further improved reconstruction quality [12]. Beyond this, multimodal frameworks have combined attention mechanisms, graph convolutional networks, and bidirectional LSTM layers to extract spatiotemporal-frequency semantic features for enhanced emotion recognition [13]. NeuroBCI systems facilitate multi-brain-to-multi-robot interactions, employing dynamic EEG feature extraction, semantic transmission frameworks, and edge computing for parallel multi-agent control [14]. Further, EEG-based word processing studies have identified theta-band activity in the left anterior temporal lobe as a key node linking distributed modality-specific networks into a multimodal semantic hub [15].

Recent studies have leveraged unsupervised semantic disentanglement to separate meaningful high-level features from neural activity, enhancing visual category decoding while improving interpretability through structured latent representations

[16]. A comprehensive review spanning EEG, functional magnetic resonance imaging, magnetoencephalography, and electrocorticography modalities has identified the absence of a universal, task-independent semantic feature, highlighting issues like inter-subject variability and multimodal integration challenges [17]. Similarly, object-based decision tasks have demonstrated notable classification accuracy using common spatial patterns and random forest classifiers for semantic category decoding [18].

These studies showcase significant strides in classification, interpretation, and semantic feature extraction using advanced techniques-from ensemble learning and deep networks to multimodal fusion and signal decomposition. However, they remain constrained by task specificity and often rely on supervised paradigms. No existing methodology provides a universal, task-independent semantic feature quantifying semantic richness in EEG data, regardless of context.

This research seeks to address this limitation by proposing a framework for extracting a universal, task-independent semantic feature from EEG signals. Building on the foundational advancements outlined above, this work introduces a novel methodology to advance EEG signal processing and semantic analysis, addressing task specificity and generalizability challenges.

3. Methodology

Extracting a universal, task-independent semantic feature from EEG signals requires a clear conceptual foundation. In this section, we define the notion of a semantic feature in EEG data, establish the criteria such a feature must satisfy, and introduce our proposed computational framework designed to achieve these objectives.

A *semantic feature* for EEG data encapsulates the high-level, contextually meaningful information encoded in neural activity, moving beyond the low-level, task-specific attributes (e.g. spectral power, temporal metrics) typically extracted. Rather than focusing narrowly on a particular frequency band or experimental condition, semantic features aim to capture the deeper cognitive or conceptual content processed by the brain, regardless of the external task or stimulus.

To qualify as a universal semantic feature, the representation must satisfy the following requirements:

1. **Task independence**: remain applicable across varying tasks, stimuli, and conditions without relying on task-specific labels or supervised models.
2. **Robustness to inter-subject variability**: consistently represent semantic content across individuals, accommodating anatomical and functional differences.
3. **Scalability and generalizability**: scale effectively to large datasets and generalize to new, unseen data with minimal performance degradation.
4. **Interpretability**: convey meaningful information that aids in understanding underlying neural processes.
5. **Compatibility with downstream analyses**: support a wide range of downstream tasks, including both supervised (e.g. classification) and unsupervised (e.g. clustering) analyses.

Modern DL and unsupervised learning techniques provide promising avenues for extracting such universal features. Unlike traditional machine learning (ML) approaches that rely on labeled datasets and are confined to specific tasks, unsupervised and self-supervised DL methods learn hierarchical representations directly from raw data, naturally aligning with our goal of task independence.

We, therefore, propose a novel computational framework that integrates three key DL components:

1. **CNNs**: capture local spatial and temporal dependencies within multichannel EEG signals.
2. **AutoEncoders**: reduce dimensionality and filter noise by learning compressed latent representations, retaining only the most salient EEG features.
3. **Transformers**: exploit self-attention mechanisms to model long-range temporal relationships and global context, extracting high-level semantic patterns independent of task conditions.

3.1. Model architecture

Our model (figure 1) is a hierarchical dual AutoEncoder designed for unsupervised EEG feature extraction. The architecture consists of two sequential AutoEncoders:

1. **CNN AutoEncoder:** the first stage applies a CNN encoder composed of two Conv1D layers with 64 and 128 filters (kernel size = 3, rectified linear unit (ReLU) activation) followed by layer normalization. This module captures local spatial–temporal dependencies in EEG signals. The output is projected into a fixed latent space of size (window_size, 128), forming an intermediate representation of the EEG segments. A CNN decoder, mirroring the encoder, later reconstructs the EEG features from this latent representation using two Conv1DTranspose layers.
2. **Transformer AutoEncoder:** the second stage processes the CNN-extracted features to capture global dependencies across the EEG sequence. The Transformer encoder consists of two layers, each with eight attention heads, an embedding size of 128, and position-wise feed-forward networks (FFNs). The Transformer decoder

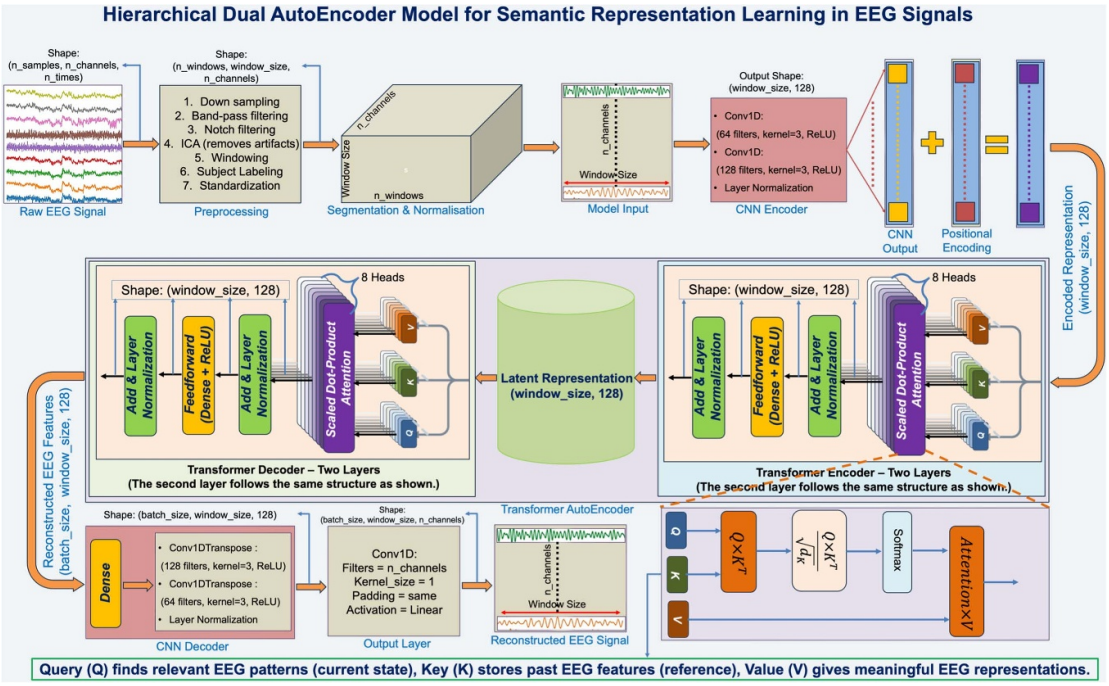


Figure 1. Overview of the proposed hierarchical dual AutoEncoder model for unsupervised EEG feature extraction. The model consists of two AutoEncoder structures: (1) CNN AutoEncoder, which extracts local temporal–spatial EEG features via a CNN encoder and reconstructs them through a CNN decoder, and (2) Transformer AutoEncoder, which captures long-range dependencies in the encoded representation using self-attention mechanisms. The EEG signals undergo preprocessing, segmentation, and normalization before being encoded by the CNN. The Transformer further processes the CNN features, producing a latent representation that retains both short-term and long-term EEG characteristics. The Transformer decoder reconstructs the EEG features, which the CNN decoder then processes to reconstruct the final EEG signals. The self-attention mechanism enables the model to capture long-range dependencies, where Q represents the current EEG state, K stores past EEG features, and V provides meaningful EEG representations.

reconstructs the EEG features using a self-attention mechanism and feed-forward layers before passing them to the CNN decoder for final signal reconstruction. Unlike standard sequence-to-sequence (Seq2Seq) models [19], our Transformer decoder does not require a cross-attention layer. This is because our model operates as an AutoEncoder, where the input and output are both EEG signals rather than different modalities or transformed sequences. In Seq2Seq models, the decoder requires cross-attention to incorporate information from an external source (e.g. translating text or generating captions). However, in our case, the Transformer decoder reconstructs the latent EEG features without needing additional information, making self-attention sufficient.

This hierarchical design allows the model to preserve both local and global dependencies while learning task-independent EEG representations. The final reconstructed EEG signals match the original input resolution, ensuring minimal loss of information.

Table 1. MI datasets with their specifications.

Attribute	BCICIV_2a [20]	BCICIV_2b [21]
Subjects	9	9
Channels	22	3
Classes	4	2
Trials/class	144	360
Trial Duration	4 s	4.5 s
Sampling Rate	250 Hz	250 Hz
Sessions	2	5
Runs	6	1
Total trials	62 208	32 400

Table 2. SSVEP datasets with their specifications.

Attribute	Lee2019_SSVEP [23]	Nakanishi2015 [22]
Subjects	54	9
Channels	62	8
Classes	4	12
Trials/class	50	15
Trial length	4 s	4.15 s
Sampling rate	1000 Hz	256 Hz
Sessions	2	1

3.2. Datasets

To validate the effectiveness of our model, we utilized a diverse collection of publicly available EEG datasets encompassing MI, steady-state visually evoked potentials (SSVEPs), and ERP, specifically P300 datasets. The details of these datasets are provided in tables 1–3.

Table 3. ERP datasets with their specifications.

Dataset	Subjects	Channels	Trials/class	Trial duration	Sampling rate	Sessions
BI2012 [24]	25	16	640 NT / 128 T	1 s	128 Hz	2
BI2013a [25–27]	24	16	3200 NT / 640 T	1 s	512 Hz	8[1]
BI2014b [28]	38	32	200 NT / 40 T	1 s	512 Hz	3
BI2015a [29]	43	32	4131 NT / 825 T	1 s	512 Hz	3
BI2015b [30]	44	32	2160 NT / 480 T	1 s	512 Hz	1
BNCI2014_008 [31, 32]	8	8	3500 NT / 700 T	1 s	256 Hz	1
BNCI2014_009 [33]	10	16	1440 NT / 288 T	0.8 s	256 Hz	3
Sosulski2019 [34–36]	13	31	7500 NT / 1500 T	1.2 s	1000 Hz	1

Note 1: for the BI2013a dataset, there are eight sessions for subjects 1–7 and one session for all other subjects.

3.3. Preprocessing

To ensure consistency across all datasets and facilitate task-independent feature extraction, we applied a standardized preprocessing pipeline to all EEG signals. This pipeline was designed to mitigate noise, improve signal quality, and standardize temporal resolution across datasets before feature extraction. The preprocessing steps were applied uniformly to all the datasets as follows:

1. **Downsampling to 128 Hz:** EEG signals were resampled to 128 Hz, ensuring a fixed temporal resolution across datasets. This standardization helps prevent discrepancies in window sizes and feature extraction.
2. **Band-pass filtering:**
   - For MI datasets, a 6–32 Hz band-pass filter was applied to retain motor-related mu and beta rhythms.
   - For ERP datasets, a 0.1–30 Hz band-pass filter was applied to preserve slow ERP components.
   - For SSVEP datasets:
     - **Lee2019_SSVEP:** a 4–14 Hz band-pass filter was applied to capture responses from the four stimulus frequencies (5.45, 6.67, 8.57, 12 Hz).
     - **Nakanishi2015:** a 8–16 Hz band-pass filter was applied to cover all twelve stimulus frequencies (9.25, 9.75, 10.25, 10.75, 11.25, 11.75, 12.25, 12.75, 13.25, 13.75, 14.25, 14.75 Hz).

   Filter design and zero-phase criterion. All band-pass filters were designed as linear-phase FIR filters with specific specifications on pass-band ripple (< 0.1 dB) and stop-band attenuation (> 40 dB). A forward-backward filtering approach (`filtfilt`) was applied offline to achieve zero-phase in the processed signals. This double-pass method cancels out any phase shifts introduced in the forward pass, ensuring no net phase distortion in the final EEG signals. We chose linear-phase FIR filters because they are straightforward to design for exact control over the frequency response and are commonly used in EEG pipelines to avoid the phase distortion that can arise with IIR filters.
3. **Notch filtering:** a linear-phase FIR notch filter at 50 Hz was similarly applied in a forward-backward manner to remove powerline interference without introducing phase shifts. Although other designs (e.g. quadratic or IIR notch filters) can also be used, FIR-based forward–backward filtering is a standard practice in offline EEG analysis due to its simplicity, stable phase characteristics, and reliable attenuation of line noise.
4. **Artifact removal:** independent component analysis was then performed to decompose the signals into maximally independent components.
5. **Windowing:** EEG signals were segmented into 1 s windows with a 50% overlap, resulting in a step size of 0.5 s. Given the fixed 128 Hz sampling rate, each window contained 128 samples per EEG channel, ensuring a consistent window size of $(128 \times n_{\text{channels}})$. This strategy enhances data availability without artificial augmentation while preserving temporal continuity, allowing each segment to retain information about preceding and succeeding windows. Overlapping windows are particularly beneficial for Transformer-based models, which require large sample sizes while maintaining the global structure of each trial.
   The window size was set to 1 s across all datasets to maintain uniformity. Most ERP datasets naturally contain 1 s trials, and MI/SSVEP datasets with longer trials (e.g. 4 s) were split into multiple 1 s windows. However, for datasets with trial lengths shorter than 1 s or non-integer durations, padding was applied to align with the segmentation framework, as shown in table 4.
   Zero-padding was applied to maintain the integrity of EEG data without introducing artificial noise. This segmentation approach ensures a standardized representation across datasets, allowing a direct comparison of extracted semantic features while preserving contextual dependencies within each trial.
6. **Subject labeling:** each segmented window was assigned a subject label to retain the identity of the subject from which the signal originated. This subject labeling is distinct from task or class

Table 4. Segmentation strategy for datasets with trial lengths shorter than 1 s or non-integer durations.

Dataset	Trial length (s)	Segmentation strategy
BNCI2014_009 (ERP)	0.8	Pad to 1 s before segmentation
Sosulski2019 (ERP)	1.2	Segment into 1 s, pad 0.2 s
Nakanishi2015 (SSVEP)	4.15	Split into 4 × 1 s, pad 0.15 s
BCICIV_2b (MI)	4.5	Split into 4 × 1 s, pad 0.5 s

labeling used in supervised learning. Since our approach is unsupervised, we do not assign task-specific labels. Instead, subject labeling is used solely to track which windows belong to which subject. This is crucial for organizing extracted latent features after training, allowing us to save and analyze features for each subject separately.

7. **Standardization:** EEG signals were z-score normalized per channel to have zero mean and unit variance, reducing inter-subject variability.

3.4. Training procedure

We train the model by minimizing the mean squared error (MSE) loss, using Adam with a learning rate of 1×10^{-4}, a batch size of 64, and up to 100 epochs. EarlyStopping (patience = 5) ensures training halts when validation loss stops improving, and ReduceLROnPlateau (factor = 0.5, patience = 3) reduces the learning rate if progress stagnates. A 90:10 train-validation split monitors generalization and informs hyperparameter tuning. Parameters like Transformer depth, batch size, or training epochs can be adjusted for datasets with different levels of complexity.

3.5. Evaluation and analysis

The evaluation of our model is conducted in two stages:

Stage 1: assessing the quality of the reconstructed EEG signals. The first evaluation phase focuses on the model's ability to reconstruct EEG signals accurately. Since the AutoEncoder structure is designed to learn a latent space that preserves the essential features of EEG signals, the reconstruction quality serves as a crucial metric. A high reconstruction error indicates that the latent representations do not effectively capture the EEG signal's structure, which would impact their usability in subsequent tasks. We use MSE between the original and reconstructed EEG signals to quantify reconstruction quality, aiming for minimal reconstruction loss. To ensure that the extracted features remained meaningful, we imposed a strict reconstruction error threshold of 1%, prompting model tuning whenever this limit was exceeded. This ensures that the Transformer AutoEncoder efficiently encodes EEG patterns into meaningful latent features.

Stage 2: evaluation of the extracted semantic features in downstream tasks. We evaluate the extracted semantic features in supervised classification and unsupervised clustering tasks once the reconstruction quality is deemed satisfactory (i.e. low MSE loss). These features are used as input to an advanced 2D CNN classifier designed to generalize across different EEG paradigms (MI, SSVEP, ERP). The architecture consists of multiple convolutional layers (64, 128, 256, 512 filters) with 3×3 kernels, batch normalization, and a spatial attention mechanism to enhance discriminative feature extraction. A global average pooling layer reduces dimensionality before classification. The final dense layer employs softmax for MI and SSVEP (multi-class classification) and sigmoid for ERP (binary classification with AUC-ROC as the evaluation metric). We apply dropout (0.5), batch normalization, and early stopping to prevent overfitting. We use focal loss for ERP datasets with imbalanced classes to improve minority class detection. Leave-one-subject-out cross-validation is applied for MI datasets due to limited subjects, while an 8-fold cross-validation strategy is used for ERP and SSVEP datasets, balancing computational feasibility and reliability. The proposed 2D CNN ensures robust classification across all paradigms, capturing spatial-temporal dependencies in the extracted semantic features while effectively handling both balanced and imbalanced datasets.

Beyond classification, we performed complementary analyses to gain deeper insights into the structure and stability of the semantic features. Clustering and dimensionality reduction techniques were employed to determine whether these features inherently form meaningful groups, while correlation analysis helped identify and remove feature redundancies. Additionally, ablation studies examined the effect of modifying Transformer complexity (e.g. number of layers, attention heads) on the quality and fidelity of the extracted features.

4. Mathematical formulation

We consider EEG signals recorded across diverse paradigms, subject populations, and experimental conditions, aiming to extract universal, task-independent semantic features using a hierarchical dual AutoEncoder model. The objective is to reconstruct EEG signals while simultaneously extracting robust, generalizable latent representations, ensuring adaptability across different datasets, recording setups, and application domains.

Let $S = \{1, 2, \ldots, N_s\}$ represent the set of subjects, and for each subject $s \in S$, EEG recordings are divided into n_s segments. These segments are denoted as $\mathcal{X}^{(s)} = \{\mathbf{X}_1^{(s)}, \mathbf{X}_2^{(s)}, \ldots, \mathbf{X}_{n_s}^{(s)}\}$, where each $\mathbf{X}_i^{(s)} \in \mathbb{R}^{T \times C}$ contains T time samples per segment and C channels. The transformation process involves two levels of encoding and decoding, structured as follows.

4.1. CNN AutoEncoder

The first stage applies a CNN AutoEncoder to learn localized representations from EEG signals. The CNN encoder transforms raw EEG segments $\mathbf{X}$ into latent feature representations $\mathbf{H}$. We first reshape $\mathbf{X} \in \mathbb{R}^{T \times C}$ to $\mathbf{X}_{\text{reshaped}} \in \mathbb{R}^{C \times T}$ and apply two 1D CNN layers along the temporal dimension:

$$\mathbf{H}^{(1)} = \text{ReLU}\left(\text{Conv1D}_{64}\left(\mathbf{X}_{\text{reshaped}}\right)\right) \tag{1}$$

$$\mathbf{H}^{(2)} = \text{ReLU}\left(\text{Conv1D}_{128}\left(\mathbf{H}^{(1)}\right)\right) \tag{2}$$

where Conv1D_{64} and Conv1D_{128} are 1D CNNs with 64 and 128 filters, kernel size 3, and same-padding. The activation function used is ReLU, which introduces non-linearity while preventing vanishing gradients in deep networks. The output is then normalized:

$$\mathbf{H}_{\text{CNN}} = \text{LayerNorm}\left(\mathbf{H}^{(2)}\right). \tag{3}$$

This feature representation is then embedded into a fixed-dimensional latent space $\mathbb{R}^{T \times d_{\text{model}}}$, where d_{model} represents the dimensionality of the latent feature space. This parameter controls the capacity of the model to encode meaningful representations, balancing expressiveness and computational efficiency. A larger d_{model} allows richer feature extraction but increases model complexity, while a smaller d_{model} enforces a more compact representation. In our analysis, we set $d_{\text{model}} = 128$ across all datasets to balance computational efficiency and feature richness. However, this value can be adjusted based on the available computational resources and the complexity of the dataset.

Next, we use a dense layer with 128 units to map $\mathbf{H}_{\text{CNN}}$ into the d_{model} dimensional space:

$$\mathbf{H}_{\text{embed}} = \text{Dense}_{128}\left(\mathbf{H}_{\text{CNN}}\right). \tag{4}$$

Positional encoding $\mathbf{E}_{\text{pos}}$ is added:

$$\mathbf{H}_{\text{PE}} = \mathbf{H}_{\text{embed}} + \mathbf{E}_{\text{pos}} \tag{5}$$

where $\mathbf{H}_{\text{PE}} \in \mathbb{R}^{T \times d_{\text{model}}}$ serves as input to the Transformer AutoEncoder.

4.2. Transformer AutoEncoder

The Transformer AutoEncoder refines the extracted local representations into global semantic features. The Transformer encoder consists of two layers, each with multi-head self-attention (MHA):

$$\mathbf{Z}^{(l)} = \text{MHA}\left(\mathbf{H}^{(l-1)}, \mathbf{H}^{(l-1)}, \mathbf{H}^{(l-1)}\right) \tag{6}$$

where $\mathbf{H}^{(0)} = \mathbf{H}_{\text{PE}}$, and the same input $\mathbf{H}^{(l-1)}$ is used as the query (Q), key (K), and value (V) [19], to capture contextual relationships within the same sequence.

The attention scores are computed using the scaled dot-product attention mechanism:

$$\text{Attention}\left(Q, K, V\right) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{7}$$

where d_k represents the dimensionality of the key vectors. The scaling factor $\frac{1}{\sqrt{d_k}}$ prevents excessively large dot-product values, which could lead to vanishing gradients during training. This scaling is particularly important when using large-dimensional embeddings, ensuring numerical stability and improving learning dynamics. In our framework, d_k is defined as d_{model}/h, where h is the number of attention heads, allowing for balanced information distribution across multiple heads.

After the self-attention operation, the output is passed through a residual connection followed by layer normalization:

$$\mathbf{U}^{(l)} = \text{LayerNorm}\left(\mathbf{H}^{(l-1)} + \mathbf{Z}^{(l)}\right). \tag{8}$$

This step ensures stable training by preventing internal covariate shift while preserving the original input information through skip connections. Next, the representation undergoes a position-wise FFN, consisting of two fully connected layers with a ReLU activation:

$$\mathbf{F}^{(l)} = \text{ReLU}\left(\mathbf{U}^{(l)}\mathbf{W}_1^{(l)} + \mathbf{b}_1^{(l)}\right)\mathbf{W}_2^{(l)} + \mathbf{b}_2^{(l)}. \tag{9}$$

A final residual connection and layer normalization step are applied:

$$\mathbf{H}^{(l)} = \text{LayerNorm}\left(\mathbf{U}^{(l)} + \mathbf{F}^{(l)}\right) \tag{10}$$

where $\mathbf{W}_1^{(l)}, \mathbf{W}_2^{(l)}, \mathbf{b}_1^{(l)}, \mathbf{b}_2^{(l)}$ are learnable parameters. After two Transformer layers, the encoder produces the latent representation of the input EEG segment:

$$\mathbf{Z} = \mathbf{H}^{(2)} \in \mathbb{R}^{T \times 128}. \tag{11}$$

This latent feature representation $\mathbf{Z}$ is then passed to the Transformer decoder, where it undergoes self-attention and feedforward processing to reconstruct high-level semantic EEG features.

4.3. Hierarchical decoding for EEG reconstruction

The CNN decoder reconstructs the original EEG signals from latent representations extracted by the Transformer AutoEncoder. First, the Transformer decoder refines $\mathbf{Z}$ into reconstructed features:

$$\mathbf{H}_{\text{DecTrans}} = \text{TransformerDecoder}(\mathbf{Z}). \quad (12)$$

A Dense layer maps these features to 128 dimensions:

$$\mathbf{H}_{\text{dense-dec}} = \text{Dense}_{128}(\mathbf{H}_{\text{DecTrans}}). \quad (13)$$

Two Conv1DTranspose layers reconstruct the original EEG sequence:

$$\mathbf{H}_{\text{dec1}} = \text{ReLU}(\text{Conv1DTranspose}_{128}(\mathbf{H}_{\text{dense-dec}})) \quad (14)$$

$$\mathbf{H}_{\text{dec2}} = \text{ReLU}(\text{Conv1DTranspose}_{64}(\mathbf{H}_{\text{dec1}})). \quad (15)$$

Applying layer normalization:

$$\mathbf{H}_{\text{dec-norm}} = \text{LayerNorm}(\mathbf{H}_{\text{dec2}}). \quad (16)$$

A final linear layer reconstructs the EEG segment:

$$\mathbf{X}' = \text{Conv1D}_C(\mathbf{H}_{\text{dec-norm}}) \in \mathbb{R}^{T\times C} \quad (17)$$

where X' is the reconstructed EEG segment, and Conv1D_C is a 1D convolutional layer with C filters, ensuring the reconstructed EEG segment $\mathbf{X}'$ matches the original input shape $\mathbb{R}^{T\times C}$.

4.4. Training objective and semantic feature extraction

The model is trained by minimizing the MSE between the original and reconstructed EEG signals:

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \left\| \mathbf{X}_i - f_4(f_3(f_2(f_1(\mathbf{X}_i)))) \right\|_{\text{F}}^2 \quad (18)$$

where N is the total number of segments and θ represents all trainable parameters, $f_1 = f_{\text{CNN-enc}}$ (CNN encoder), $f_2 = f_{\text{Transformer-enc}}$ (Transformer encoder), $f_3 = f_{\text{Transformer-dec}}$ (Transformer decoder), and $f_4 = f_{\text{CNN-dec}}$ (CNN decoder).

After training, we extract semantic features as follows. For each segment $\mathbf{X}_i^{(s)}$, the Transformer encoder produces:

$$\mathbf{Z}_i^{(s)} = f_{\text{Transformer-enc}}\left(\mathbf{X}_i^{(s)}\right) \in \mathbb{R}^{T\times 128}. \quad (19)$$

Temporal averaging yields a fixed-size feature vector for each segment:

$$\mathbf{z}_i^{(s)} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}_i^{(s)}(t,:) \in \mathbb{R}^{128}. \quad (20)$$

Averaging over all segments of subject s gives a subject-level semantic feature vector:

$$\mathbf{z}^{(s)} = \frac{1}{n_s}\sum_{i=1}^{n_s} \mathbf{z}_i^{(s)} \in \mathbb{R}^{128}. \quad (21)$$

Normalizing $\mathbf{z}^{(s)}$ to unit norm produces:

$$\tilde{\mathbf{z}}^{(s)} = \frac{\mathbf{z}^{(s)}}{\|\mathbf{z}^{(s)}\|_2}. \quad (22)$$

The resulting $\tilde{\mathbf{z}}^{(s)}$ is the subject-level semantic feature vector, representing universal, task-independent information.

5. Results

This section presents the outcomes of our universal, task-independent semantic feature extraction framework across multiple EEG paradigms-MI, SSVEP, and ERP (P300)-summarized in tables 1–3. Our method is implemented with PyTorch 2.0+ in Python 3.11 using an NVIDIA GeForce RTX 4090 with CUDA 12.0.

To analyze the model's convergence behavior across different EEG paradigms, we present the learning dynamics for three representative datasets: BCICIV_2b (MI), BI2015a (ERP), and Lee2019_SSVEP (SSVEP). Figures 2–4 illustrate the training and validation loss curves, alongside the learning rate adjustments over training epochs. These plots demonstrate stable convergence and effective generalization across all paradigms, with the learning rate adapting dynamically to improve training efficiency. The consistent downward trend of both training and validation losses across all datasets highlights the robustness of the proposed framework in learning meaningful EEG representations.

Next, we assess the quality of EEG signal reconstruction across different paradigms. Figure 5 compares original and reconstructed EEG signals, this time for three other representative datasets: BCICIV_2b (MI), BNCI2014_008 (ERP), and Nakanishi2015 (SSVEP). The close alignment between original and reconstructed signals and low reconstruction errors confirm that the extracted latent semantic features effectively preserve essential EEG characteristics while minimizing noise.

We conducted a detailed correlation analysis for each subject to further evaluate the extracted latent semantic features. The correlation heatmaps for BCICIV_2a MI dataset, as shown in figure 6, reveal subject-specific structural patterns and provide insights into the consistency and diversity of the extracted features across individuals. These patterns underscore the model's ability to balance universal feature representation with subject-specific nuances, paving the way for robust and adaptable downstream analyses.

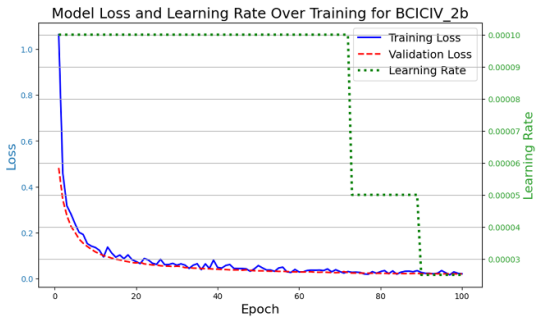


Figure 2. Training and validation loss curves alongside the learning rate schedule for the BCICIV_2b MI dataset. The model exhibits stable convergence with steadily decreasing loss, indicating effective feature learning and generalization.

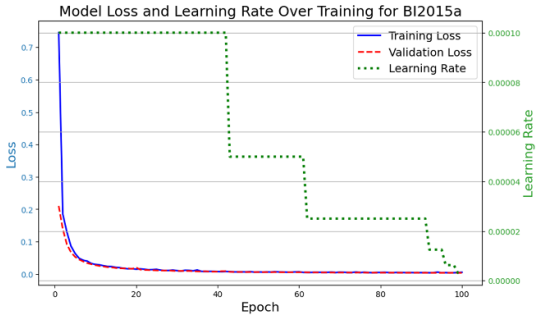


Figure 3. Training and validation loss curves with learning rate adjustments for the BI2015a ERP dataset. The smooth convergence and gradual decrease in learning rate highlight the model's adaptability and robust generalization to ERP signals.

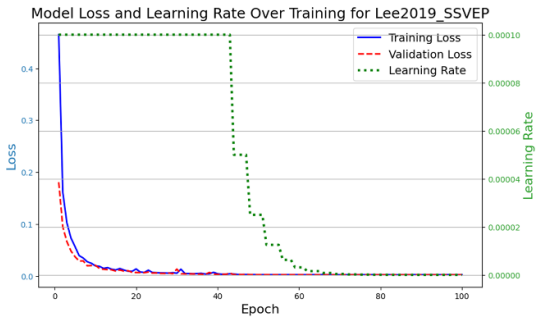


Figure 4. Training and validation loss curves with learning rate dynamics for the Lee2019_SSVEP dataset. The model efficiently minimizes loss while adjusting the learning rate, demonstrating effective training on SSVEPs.

We evaluated the semantic features in supervised classification tasks for MI, SSVEP, and ERP datasets. Although detailed comparisons with state-of-the-art methods are discussed later, tables 5–8 show that our approach performs favorably. Accuracy is reported for MI and SSVEP (balanced classes) and AUC for ERP (P300) datasets (class imbalance). Subsequent sections provide interpretations and in-depth discussions of these findings. The acronyms (e.g. FBCSP, EEGNet) used in tables 5–8 refer to models as originally presented in their respective studies, with details available in the cited references.

To comprehensively evaluate the classification performance on the BCICIV_2a dataset, we analyzed the ROC curves for each subject and aggregated the results to highlight overall performance trends. For brevity, figure 7 presents the aggregate ROC curve over all subjects, summarizing the discriminative capability of the model across the four MI classes, with high AUC values indicating robust classification performance.

Finally, to further evaluate the structure and separability of the extracted features, we applied t-distributed stochastic neighbor embedding (t-SNE) to visualize both raw EEG data and the latent feature representations across different paradigms. Figure 8 presents t-SNE embeddings of EEG data for three representative datasets: BCICIV_2b (MI), Sosulski2019 (ERP), and Nakanishi2015 (SSVEP). The left column displays the raw EEG signals, while the right column represents the latent features learned by our proposed framework. A key observation is that raw EEG data does not exhibit clear clusters, as subject-specific distributions are significantly overlapped. This confirms that EEG signals in their raw form are highly variable and noisy, making direct classification challenging. In contrast, the latent features form well-separated clusters, indicating that our feature extraction method successfully captures subject-specific representations while reducing noise.

These results confirm that the extracted semantic features are well-structured and more informative than raw EEG data, further explaining the improved classification performance across multiple EEG paradigms.

6. Discussion

This section delves into the detailed interpretation of our findings, contextualizing the results presented earlier and exploring their implications for EEG-based semantic feature extraction.

6.1. Convergence behavior and generalization

The learning curves across different EEG paradigms confirm the model's ability to generalize effectively while maintaining stable convergence. Figures 2–4 illustrate the training and validation loss trends, as well as the learning rate dynamics, for three representative datasets.

Across all three paradigms, the training loss steadily decreases, with validation loss closely following a similar trend, indicating effective feature learning and strong generalization capabilities. The learning rate adjustments further illustrate how the model dynamically adapts during training, reducing the step size at critical points to refine optimization and prevent overfitting. These adjustments are particularly

Table 5. Comparisons with state-of-the-art methods on dataset BCICIV_2a (Accuracy).

Methods	S01	S02	S03	S04	S05	S06	S07	S08	S09	Average
FBCSP [37]	76.00	56.50	81.25	61.00	55.00	45.25	82.75	81.25	70.75	67.75
ConvNet [38]	76.39	55.21	89.24	74.65	56.94	54.17	**92.71**	77.08	76.39	72.53
EEGNet [39]	85.76	61.46	88.54	67.01	55.90	52.08	89.58	83.33	86.81	74.50
C2CM [40]	87.50	65.28	90.28	66.67	62.50	45.49	89.58	83.33	79.51	74.46
FBCNet [41]	85.42	60.42	90.63	76.39	**74.31**	53.82	84.38	79.51	80.90	76.20
DRDA [42]	83.19	55.14	87.43	75.28	62.29	57.15	86.18	83.61	82.00	74.74
Conformer [43]	88.19	61.46	93.40	78.13	52.08	65.28	92.36	88.19	**88.89**	78.66
This study	**89.65**	**71.69**	**95.16**	**86.19**	65.09	**76.17**	89.99	**92.40**	85.15	**83.50**
Average accuracy	84.01	60.90	89.49	73.17	60.51	56.18	88.44	83.59	81.30	—

Table 6. Comparisons with state-of-the-art methods on dataset BCICIV_2b (Accuracy).

Methods	S01	S02	S03	S04	S05	S06	S07	S08	S09	Average
FBCSP [37]	70.00	60.36	60.94	97.50	93.12	80.63	78.13	92.50	86.88	80.00
ConvNet [38]	76.56	50.00	51.56	96.88	93.13	85.31	83.75	91.56	85.62	79.37
EEGNet [39]	75.94	57.64	58.43	98.13	81.25	88.75	84.06	93.44	89.69	80.48
DRDA [42]	81.37	62.86	63.63	95.94	**93.56**	88.19	85.00	**95.25**	90.00	83.98
Conformer [43]	82.50	65.71	63.75	**98.44**	86.56	**90.31**	**87.81**	94.38	**92.19**	84.63
This study	**89.5**	**78.5**	**71.5**	96.67	87.86	85.25	83.25	87.05	84.00	**84.84**
Average	79.31	62.51	61.64	97.26	89.25	86.41	83.67	92.36	88.06	—

Table 7. Comparisons with state-of-the-art methods on SSVEP datasets (Accuracy).

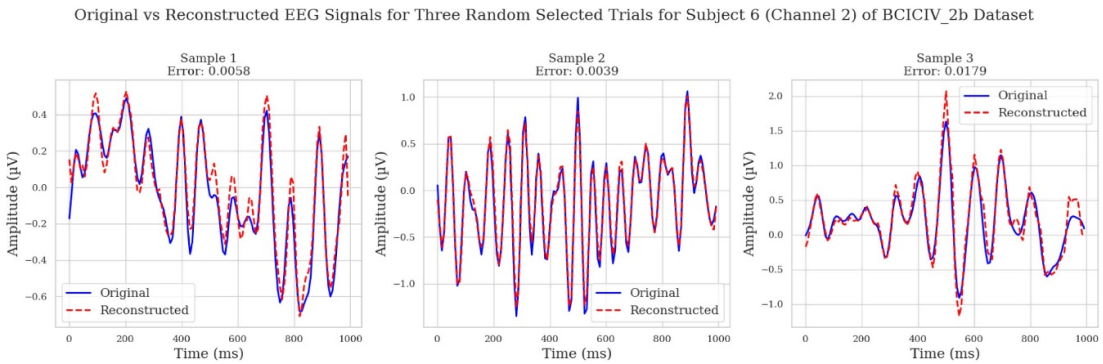
Methods	Lee2019_SSVEP	Nakanishi2015
CCA [44]	90.97	92.53
EEGITNet [45]	86.84	80.86
EEGNeX [46]	93.81	82.65
EEGNet [39]	64.43	44.14
MDM [47]	75.38	78.77
Barachant *et al* [48]	89.44	87.22
ConvNet [38]	69.36	57.47
TRCA [49]	97.78	99.20
This study	**98.41**	**99.66**

Table 8. Comparisons with state-of-the-art methods on P300 datasets (AUC).

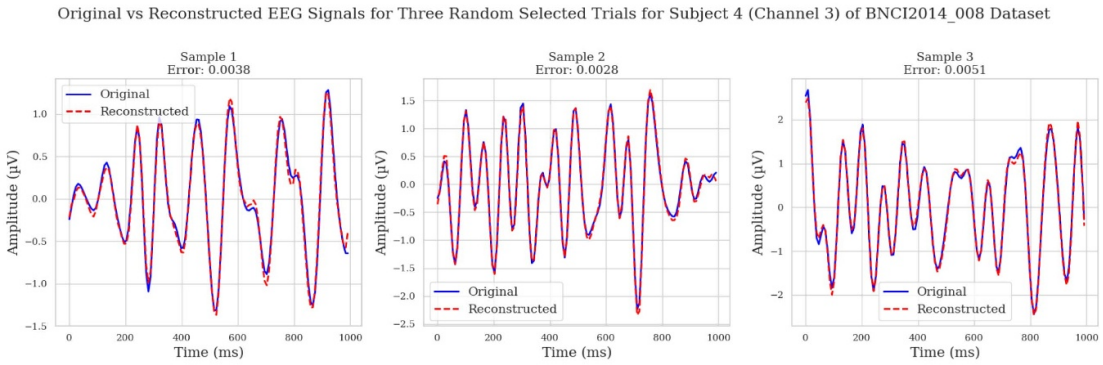
Methods	BNCI2014-008	BNCI2014-009	BI2012	BI2013a	BI2014b	BI2015a	BI2015b	Sosulski2019	Average
EEGITNet [45]	86.00	92.21	89.65	90.01	86.27	90.71	83.33	88.82	88.37
EEGNeX [46]	83.86	90.58	88.22	88.62	83.87	87.62	81.60	86.18	86.39
EEGNet [39]	85.91	91.37	87.13	85.40	80.14	86.80	86.63	87.14	86.31
ERPCov+MDM [26]	74.30	81.16	82.90	82.02	71.62	77.52	72.07	68.17	76.22
ERPCov(svdn4)+ MDM [26]	75.42	84.52	79.02	82.07	76.48	77.92	77.09	70.63	77.89
ConvNet [38]	81.07	85.12	77.06	74.50	63.75	59.56	73.20	78.35	74.07
LDA [38]	82.24	64.03	76.74	76.60	73.02	76.02	77.74	67.49	74.23
Barachant *et al* [50]	77.62	92.04	88.22	**97.09**	88.58	92.57	83.48	86.07	88.21
Chevallier *et al* [51]	85.61	**93.15**	90.99	92.71	**91.88**	**93.05**	84.56	**98.44**	91.30
This study	**96.74**	91.95	**92.86**	90.32	89.93	91.43	**95.12**	86.07	**91.80**

evident in the BI2015a and Lee2019_SSVEP datasets, where the learning rate follows a step-wise decay strategy, improving convergence stability.

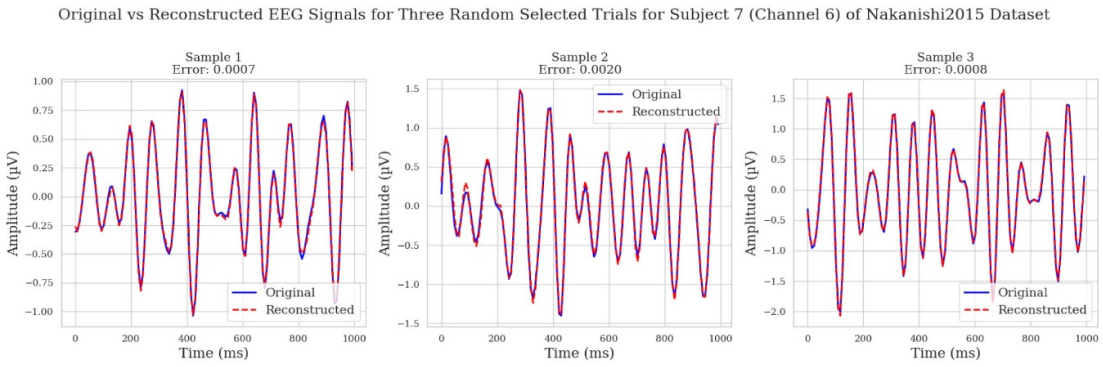
Notably, the smooth and progressive reduction in loss across all datasets confirms the proposed framework's robustness in learning task-independent semantic features, ensuring adaptability to different EEG paradigms. These findings emphasize the model's ability to extract meaningful representations from diverse EEG signals, making it applicable across a wide range of neurophysiological studies and BCI applications.



(a) MI Representative - BCICIV_2b



(b) ERP Representative - BNCI2014_008



(c) SSVEP Representative - Nakanishi2015

Figure 5. Comparison of original and reconstructed EEG signals for three randomly selected trials from different datasets. (a) BCICIV_2b (MI): Subject 6, Channel 2. (b) BNCI2014_008 (ERP): Subject 4, Channel 3. (c) Nakanishi2015 (SSVEP): Subject 7, Channel 6. The strong alignment between original (solid blue) and reconstructed (dashed red) signals, along with low reconstruction errors, demonstrates the effectiveness of the latent semantic features in capturing and preserving meaningful EEG information across paradigms.

6.2. Reconstruction performance

The quality of EEG signal reconstruction across different paradigms highlights the fidelity and utility of the extracted latent semantic features. Figure 5 showcases representative examples from three paradigms. While these datasets serve as exemplars, we observed similar reconstruction trends across all 12 datasets, demonstrating the robustness of our framework.

Among the paradigms, MI proved to be the most challenging, followed by ERP, with SSVEP being the easiest to reconstruct. This trend is evident in the reconstruction error distributions across datasets, as MI signals exhibit higher intra-subject variability and are more complex, making their latent feature extraction and reconstruction more difficult. While still containing transient components, ERP signals were more structured and showed improved reconstruction fidelity. Finally, SSVEP signals, characterized by periodic and highly structured responses to visual stimuli, exhibited the highest reconstruction quality.

A lower reconstruction error indicates higher-quality extracted features, confirming that the AutoEncoder-based model effectively captures and preserves the critical information in EEG signals.

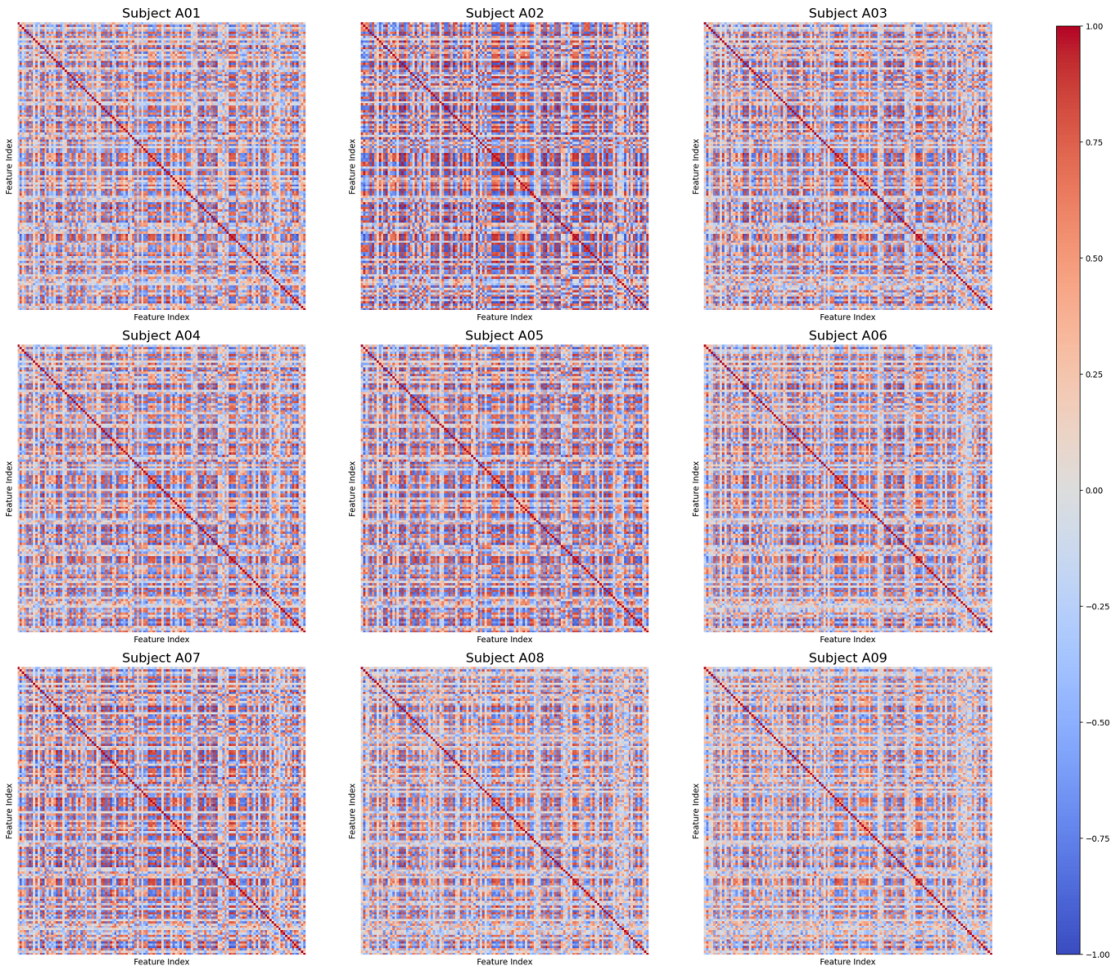


Figure 6. Correlation heatmaps of latent features for each subject of the BCICIV_2a MI Dataset: this figure presents the correlation matrices of latent features extracted from EEG signals for all subjects. The diagonal represents perfect correlation (value = 1), while off-diagonal values indicate varying degrees of correlation between different features. The patterns suggest underlying structure and dependencies in the learned feature representations.

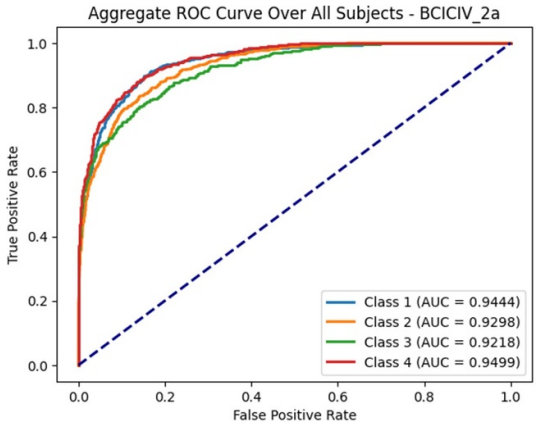


Figure 7. Aggregate ROC curve over all subjects for BCICIV_2a MI dataset: the plot presents the ROC curves for all four classes, showing the model's classification performance across subjects. The AUC values indicate high discriminative ability, with all classes achieving values above 0.92.

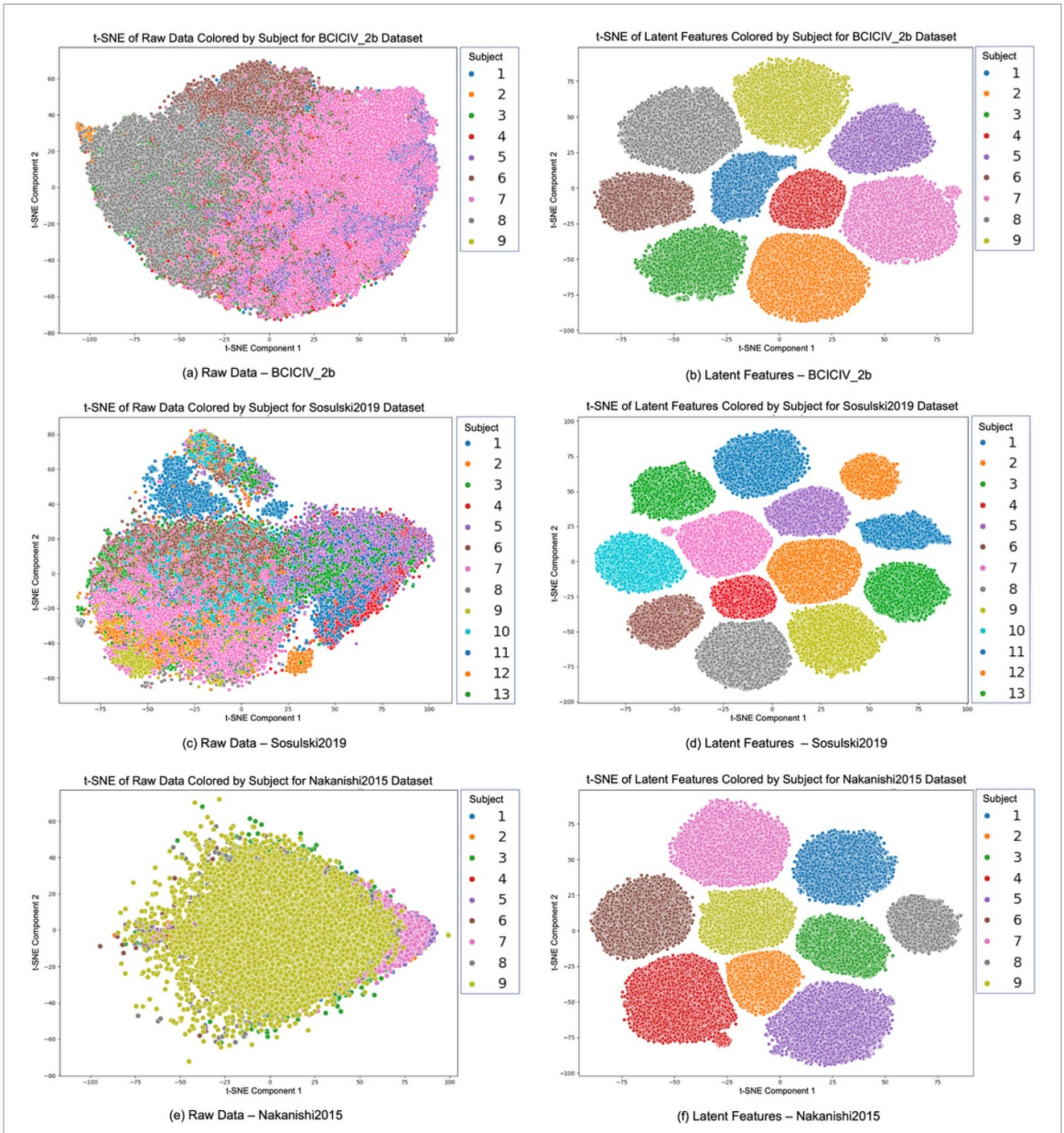


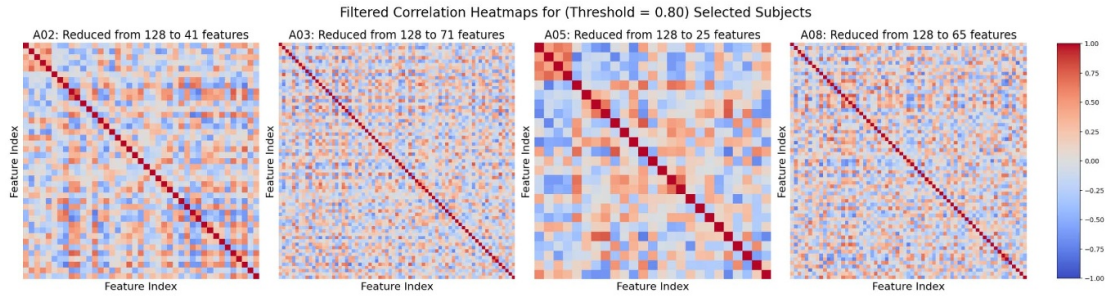
Figure 8. t-SNE visualization of EEG data before and after feature extraction for three representative paradigms: BCICIV_2b (MI), Sosulski2019(ERP), and Nakanishi2015(SSVEP). The left column shows raw EEG data, while the right column displays extracted latent features. The clustering in the right column demonstrates that latent features capture subject-specific structures more effectively than raw EEG data.

Notably, all 12 datasets had reconstruction errors well below the threshold (1%).

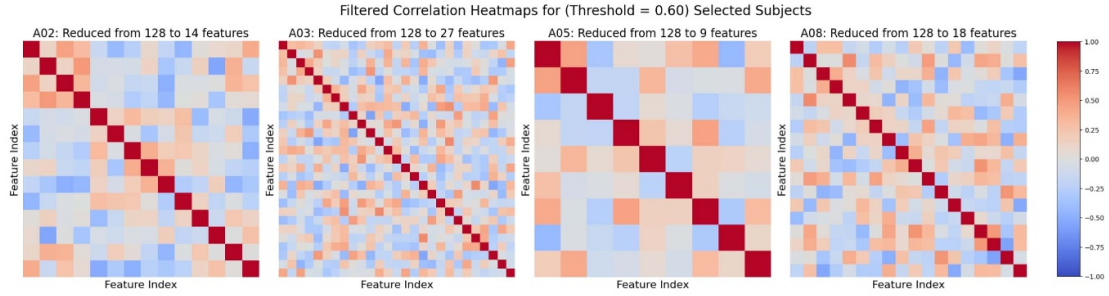
The best reconstruction performance was observed for the Nakanishi2015 (SSVEP) dataset, with an overall reconstruction error of 0.0021, reaffirming the ease of reconstructing EEG signals with quasi-periodic responses to rhythmic visual stimulation. In contrast, the BCICIV_2a (MI) dataset exhibited the highest reconstruction error, at 0.0094, aligning with the more significant challenge of MI signals. Despite this, the reconstruction quality for all datasets remained within an acceptable range, demonstrating the adaptability of our framework to different EEG paradigms.

6.3. Supervised classification performance and comparison with state-of-the-art methods

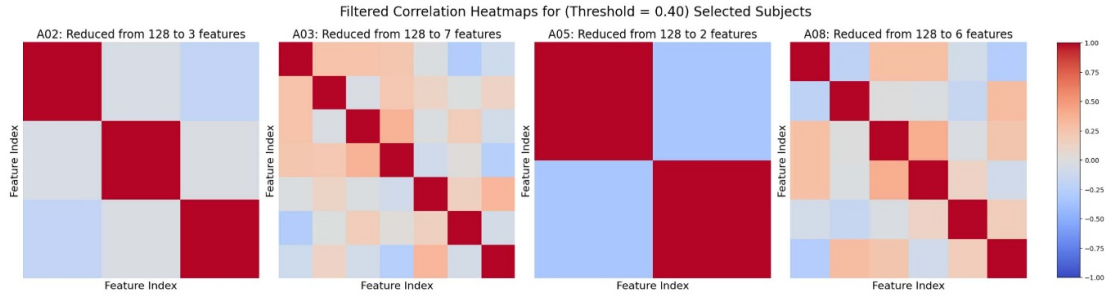
We assessed the classification performance of our extracted semantic features across all datasets, comparing results against state-of-the-art approaches (tables 5–8). For MI datasets (BCICIV_2a and BCICIV_2b), our framework achieved top accuracies of 83.50% and 84.84%, respectively, outperforming leading approaches such as Conformer [43], FBCNet [41], DRDA [42], and demonstrating robustness and adaptability to subject variability. On SSVEP datasets (Lee2019-SSVEP and Nakanishi2015), we recorded accuracies of 98.41% and 99.66%, surpassing TRCA [49] (previous best) and confirming our model's



(a) Threshold = 0.80



(b) Threshold = 0.60



(c) Threshold = 0.40

Figure 9. Filtered correlation heatmaps for selected subjects of BCICIV_2a dataset at different thresholds (0.80, 0.60, 0.40): these heatmaps illustrate the effect of varying correlation thresholds on feature selection. Higher thresholds (0.80) retain more features with strong dependencies, while lower thresholds (0.40) result in fewer retained features, capturing only the most distinct information. Lower-performing subjects (A02, A05) consistently retain fewer features across all thresholds, indicating higher redundancy, whereas higher-performing subjects (A03, A08) retain more features, suggesting a richer and more informative feature set.

suitability for steady-state stimulus decoding. For ERP (P300) datasets, we attained the highest average AUC (91.80%) across eight datasets, exceeding results from [50, 51], and setting new performance benchmarks in handling class-imbalanced conditions. Collectively, these results attest to the generality and resilience of our semantic feature extraction, capturing meaningful neural patterns that apply to diverse tasks. They highlight the approach's scalability, versatility, and potential real-world utility, while establishing a new benchmark for EEG-based classification and advancing both theoretical and practical aspects of EEG analysis.

In addition to achieving high accuracy and AUC values, the ROC analysis provides deeper insights into the discriminative power of the extracted semantic features. The aggregate ROC curve for the BCICIV_2a MI dataset, presented in figure 7, consolidates the class-wise performance across all subjects, yielding high AUC values for all four classes (ranging from 0.9218 to 0.9499). These results emphasize the model's robust ability to distinguish between different classes effectively, further reinforcing its generalization and adaptability in diverse classification scenarios.

6.4. Semantic feature correlation and performance analysis

Each participant exhibits a unique correlation matrix, reflecting individual neural dynamics as shown in figure 6. High positive correlations indicate feature pairs that capture similar neural processes, reinforcing the idea of task-independent representations. The presence of both complementary and overlapping features suggests that balancing redundancy reduction with feature

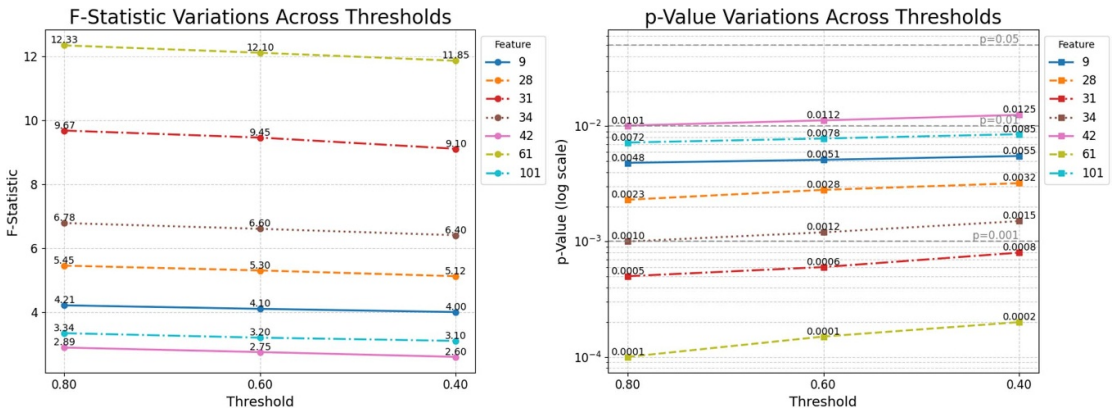


Figure 10. F-statistic and p-value variations across different correlation thresholds (0.80, 0.60, and 0.40) for selected features from Subject A03 in BCICIV_2a MI dataset. The left plot shows how the F-statistic changes with different thresholds, where a higher F-statistic indicates stronger discriminative power. The right plot presents the p-values on a logarithmic scale, illustrating statistical significance trends. Dashed reference lines at $p = 0.05$, $p = 0.01$, and $p = 0.001$ highlight significance thresholds. Features with higher F-statistics generally exhibit lower p-values, confirming their relevance for classification.

diversity enhances feature quality. These results confirm that the proposed framework extracts a mix of universal and subject-specific semantic features, leading to robust and meaningful representations applicable to various EEG-related tasks.

Understanding the correlation among extracted semantic features provides valuable insights into their structure, diversity, and relevance to classification performance. Figure 9 presents filtered correlation heatmaps for selected subjects across different thresholds (0.80, 0.60, and 0.40), illustrating how feature selection is influenced by varying correlation levels.

Furthermore, we investigated whether subject-level classification performance correlates with the diversity and quality of extracted semantic features. As seen in figure 9, lower-performing subjects (e.g. A02 and A05) consistently retained fewer features across all thresholds, indicating higher redundancy in their feature sets. In contrast, high-performing subjects (e.g. A03 and A08) retained a greater number of features, suggesting a richer and more informative feature space.

To systematically assess the impact of different correlation thresholds, we analyzed feature retention at three levels:

1. **At 0.80,** a stricter threshold retains highly correlated features, leading to a larger number of preserved features. Subjects A03 and A08 retained 71 and 65 features, respectively, while subjects A02 and A05 retained 41 and 25 features.
2. **At 0.60,** a more relaxed threshold reduces redundancy further, leading to fewer retained features while maintaining essential feature diversity. Subjects A03 and A08 retained 27 and 18 features, whereas subjects A02 and A05 retained 14 and 9 features, respectively.
3. **At 0.40,** an even lower threshold removes most redundant features, retaining only the most distinct ones. At this level, high-performing subjects still retained relatively more informative features (7 and 6 features for A03 and A08), while lower-performing subjects were left with 3 and 2 features, indicating that their feature sets had high redundancy and limited discriminative power.

To further validate the discriminative power of the retained features, we conducted an analysis of variance test on the final 7 features retained at the lowest threshold (0.40) for Subject A03. Each feature was examined to determine whether its values differed significantly among the class labels (the factors in the analysis). Figure 10 presents the variations in F-statistic and p-values across thresholds, illustrating how feature significance evolves with different levels of correlation filtering. The results confirm that all 7 selected features remained statistically significant ($p < 0.05$) across the thresholds, reaffirming their importance in distinguishing between classes. Notably, features with higher F-statistics consistently exhibited lower p-values, reinforcing their strong discriminative capability. These findings suggest that while varying correlation thresholds affect the number of retained features, the core discriminative power of these final selected features remains stable, ensuring their robustness for classification.

These findings suggest that selective feature retention enhances classification accuracy by preserving only the most relevant and non-redundant semantic features. While the current threshold selection was exploratory, future research could focus on developing a systematic, data-driven

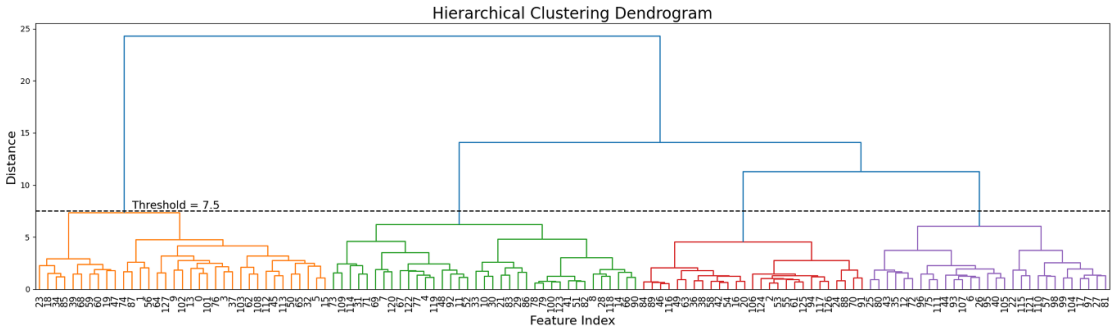


Figure 11. Hierarchical clustering dendrogram of semantic features. A distance threshold (7.5) reveals four clusters corresponding to the BCICIV_2a MI dataset classes, highlighting the features' ability to capture class-specific patterns without supervision.

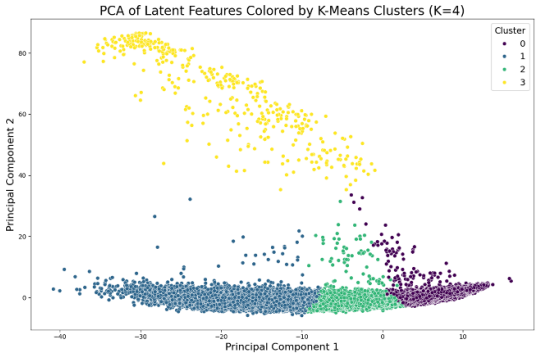


Figure 12. PCA of latent semantic features, grouped by K-means clustering ($k = 4$). The clusters demonstrate the natural grouping of features into distinct categories, corresponding to shared neural patterns across the BCICIV_2a dataset classes, underscoring the model's ability to capture task-relevant information.

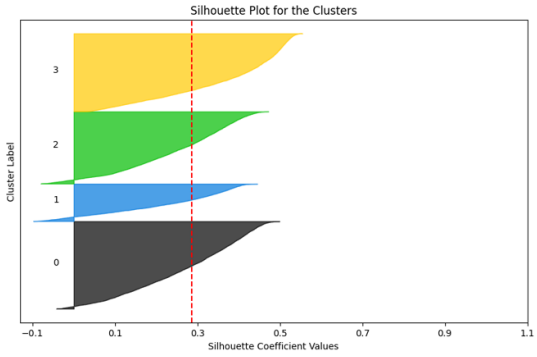


Figure 13. Silhouette plot confirms strong cluster cohesion and separation.

approach to optimize threshold selection for maximizing classification performance. By identifying and leveraging the most informative semantic features, the proposed framework enables personalized, adaptive EEG-based applications, paving the way for more interpretable and robust neural representations.

6.5. Clustering analysis of extracted semantic features

The clustering properties of the extracted semantic features provide further insights into their structure and interpretability. Hierarchical clustering (figure 11) demonstrated the ability of the extracted features to naturally separate different MI classes without explicit supervision, reinforcing their class-discriminative nature in BCICIV_2a dataset. Similarly, principal component analysis (PCA) and K-means clustering (figure 12) highlighted the inter-subject variability while maintaining task-specific organization, further confirming the robustness of the extracted feature space. We computed silhouette scores for the extracted semantic features to validate the cluster quality further, confirming strong cohesion and separation as shown in figure 13.

Beyond these clustering approaches, t-SNE visualization (figure 8) provided an intuitive representation of the separability between raw EEG signals and extracted semantic features. A key observation from t-SNE analysis is that while raw EEG data exhibits substantial overlap and lacks clear structure, the latent feature space forms well-separated clusters, particularly in high-performing datasets. This suggests that the extracted features effectively capture subject-specific patterns while reducing noise, addressing a well-known limitation of EEG signals-high variability and low signal-to-noise ratio.

Moreover, the improvement in clustering across different EEG paradigms highlights the generalizability of our approach. The formation of more compact clusters in the latent feature space indicates that the feature extraction process successfully maps the data into a more structured representation, making it inherently more suitable for downstream tasks such as classification and subject recognition.

Additionally, while PCA provides a linear projection, t-SNE is particularly useful for capturing complex, non-linear relationships. The contrast between PCA and t-SNE results further confirms that the extracted semantic features preserve essential neural dynamics while improving separability. This reinforces the effectiveness of our method in transforming high-dimensional, noisy EEG data into

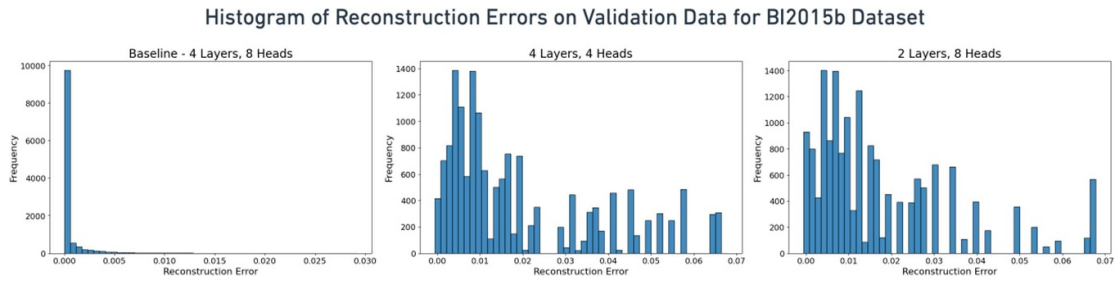


Figure 14. Ablation analysis of reconstruction error distributions for different Transformer configurations on the BI2015b dataset. The baseline (4 layers, 8 heads) achieves the lowest error variance, whereas reducing attention heads or layers increases variance, confirming the baseline's near-optimal performance.

a well-structured latent space, ultimately benefiting both supervised and unsupervised learning tasks.

The clustering results suggest that our framework does more than just enhance classification performance-it also provides a well-organized, interpretable feature space that could be leveraged for other neurophysiological analyses, including anomaly detection, cognitive state decoding, and adaptive BCIs.

6.6. Ablation analysis

To assess the impact of varying the Transformer configuration, we conducted an ablation study on the BI2015b dataset, an ERP representative comprising 44 subjects with 2160 NT / 480 T trials per class and 116 160 total trials, of which 10% (11 616 trials) were used for evaluation. Figure 14 illustrates the reconstruction error distributions for different model configurations, highlighting the effect of altering the number of layers and attention heads.

Reducing the number of attention heads from eight to four led to increased error variance, suggesting that fewer heads limit the model's capacity to capture diverse feature dependencies. Similarly, decreasing the layer count from four to two resulted in broader error distributions, emphasizing the necessity of deeper architectures for effectively modeling hierarchical representations. The baseline configuration (four layers, eight heads) demonstrated the lowest variance in reconstruction errors, confirming its near-optimal balance between model complexity and performance. These results reinforce that our chosen architecture effectively captures semantic EEG features while minimizing reconstruction errors.

6.7. Limitations and future directions

Although the proposed methodology demonstrates strong performance, several limitations must be acknowledged. Model complexity and high computational costs pose challenges for real-time or resource-constrained applications. The data-hungry nature of Transformers presents difficulties when dealing with typically small EEG datasets, raising the risk of overfitting and limited generalization. While data augmentation can address data scarcity, excessive augmentation risks diluting meaningful patterns. Furthermore, training deep models on constrained EEG datasets can lead to optimization issues, and the neurophysiological interpretation of the extracted semantic features remains unclear.

To overcome these constraints and advance the methodology, future work should focus on optimizing model configurations to maintain high accuracy while reducing complexity. EEG-specific data augmentation strategies can improve robustness without introducing noise or redundancy. Transfer learning from larger EEG datasets may enhance performance on smaller, domain-specific datasets. Additionally, neurophysiological validations are needed to establish clear links between the extracted features and underlying brain processes, improving interpretability and practical relevance. Efforts to streamline computational efficiency-through lighter architectures, pruning, or quantization-could make the approach more accessible. Finally, a theoretical framework is needed to better understand the semantic features' role and utility while systematically balancing model complexity, data size, and computational overhead will ensure that solutions remain both practical and high-performing.

7. Conclusion

We presented a universal, task-independent framework for extracting semantic features from EEG signals, bridging the gap between conventional feature extraction and task-specific methods. By integrating CNNs, AutoEncoders, and Transformers, our approach captures both spatial-temporal details and higher-level semantic representations. Evaluations across diverse EEG paradigms show that it surpasses state-of-the-art methods and yields generalizable features suitable for various research and applied domains.

Addressing high computational costs and improving feature interpretability will enhance scalability and usability. Future efforts will optimize model configurations, employ transfer learning, develop targeted augmentation strategies, and

pursue neurophysiological validations to deepen understanding of the extracted features and their neural underpinnings.

Data availability statement

No new data were created or analysed in this study.

ORCID iDs

Hossein Ahmadi https://orcid.org/0000-0002-3650-0280
Luca Mesin https://orcid.org/0000-0002-8239-2348

References

[1] Schomer D L and Lopes da Silva F H 2011 *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications and Related Fields* 6th edn (Lippincott Williams & Wilkins)
[2] Michel C M and Koenig T 2018 EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review *NeuroImage* **180** 577–93
[3] Ahmadi H and Mesin L 2023 Enhancing motor imagery electroencephalography classification with a correlation-optimized weighted stacking ensemble model *Electronics* **13** 1033
[4] Ahmadi H and Mesin L 2024 Enhancing MI EEG signal classification with a novel weighted and stacked adaptive integrated ensemble model: a multi-dataset approach *IEEE Access* **12** 103626–46
[5] Zheng W *et al* 2021 Predicting neurological outcome in comatose patients after cardiac arrest with multiscale deep neural networks *Resuscitation* **169** 86–94
[6] Ahmadi H, Kuhestani A and Mesin L 2024 Adversarial neural network training for secure and robust brain-to-brain communication *IEEE Access* **12** 39450–69
[7] Ahmadi H, Kuhestani A, Keshavarzi M and Mesin L 2025 Securing brain-to-brain communication channels using adversarial training on SSVEP EEG *IEEE Access* **13** 14358–78
[8] Fahimi Hnazaee M, Khachatryan E and Van Hulle M M 2018 Semantic features reveal different networks during word processing: an EEG source localization study *Front. Hum. Neurosci.* **12** 503
[9] Ravi Kumar M and Srinivasa Rao Y 2019 Epileptic seizures classification in EEG signal based on semantic features and variational mode decomposition *Cluster Comput.* **22** 13521–31
[10] Sartori G, Polezzi D, Mameli F and Lombardi L 2005 Feature type effects in semantic memory: an event related potentials study *Neurosci. Lett.* **390** 139–44
[11] Zeng H, Xia N, Qian D, Hattori M, Wang C and Kong W 2023 DM-RE2I: a framework based on diffusion model for the reconstruction from EEG to image *Biomed. Signal Process. Control* **86** 105125
[12] Zeng H, Xia N, Tao M, Pan D, Zheng H, Wang C, Xu F, Zakaria W and Dai G 2023 DCAE: a dual conditional autoencoder framework for the reconstruction from EEG into image *Biomed. Signal Process. Control* **81** 104440
[13] Chen J, Liu Y, Xue W, Hu K and Lin W 2022 Multimodal EEG emotion recognition based on the attention recurrent graph convolutional network *Information* **13** 550
[14] Ouyang J, Wu M, Li X, Deng H, Jin Z and Wu Di 2024 NeuroBCI: multi-brain to multi-robot interaction through EEG-adaptive neural networks and semantic communications *IEEE Trans. Mob. Comput.* **23** 14622–37
[15] van Ackeren M J and Rueschemeyer S-A 2014 Cross-modal integration of lexical-semantic features during word processing: evidence from oscillatory dynamics during EEG *PLoS One* **9** e101042
[16] Zhou Q *et al* 2025 Interpretable visual neural decoding with unsupervised semantic disentanglement *Mach. Intell. Res.* (https://doi.org/10.1007/s11633-023-1484-y)
[17] Rybář M and Daly I 2022 Neural decoding of semantic concepts: a systematic literature review *J. Neural Eng.* **19** 021002
[18] Rekrut M *et al* 2020 Decoding semantic categories from EEG activity in object-based decision tasks *Proc. 2020 8th Int. Winter Conf. on Brain-Computer Interface (BCI), Gangwon, Korea (South)* pp 1–7
[19] Vaswani A *et al* 2017 Attention is all you Need (arXiv:1706.03762)
[20] Tangermann M *et al* Review of the BCI competition IV 2012 *Front. Neurosci.* **6** 21084
[21] Leeb R, Lee F, Keinrath C, Scherer R, Bischof H and Pfurtscheller G 2007 Brain-computer communication: motivation, aim and impact of exploring a virtual apartment *IEEE Trans. Neural Syst. Rehab. Eng.* **15** 473–82
[22] Nakanishi M, Wang Y, Wang Y-T and Jung T-P 2015 A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials *PLoS One* **10** e0140703
[23] Lee M, Kwon O-Y, Kim Y-J, Kim H-K, Lee Y-E, Williamson J, Fazli S and Lee S-W 2019 EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy *GigaScience* 8 giz002
[24] Van Veen G *et al* 2019 Building brain invaders: EEG data of an experimental validation (arXiv:1905.05182)
[25] Vaineau E *et al* 2019 Brain invaders adaptive versus non-adaptive P300 brain-computer interface dataset (arXiv:1904.09111)
[26] Barachant A and Congedo M 2014 A Plug&Play P300 BCI using information geometry (arXiv:1409.0107)
[27] Congedo M *et al* 2011 Brain Invaders': a prototype of an open-source P300-based video game working with the OpenViBE platform *Proc. IBCI Conf.* (*Graz, Austria*) pp 280–3
[28] Korczowski L *et al* 2019 Brain invaders solo versus collaboration: multi-user P300-based brain-computer interface dataset (BI2014b) (available at: https://hal.archives-ouvertes.fr/hal-02173958)
[29] Korczowski L *et al* 2019 Brain Invaders calibration-less P300-based BCI with modulation of flash duration dataset (BI2015a) (available at: https://hal.archives-ouvertes.fr/hal-02172347)
[30] Korczowski L *et al* 2019 Brain invaders cooperative versus competitive: multi-user P300-based brain-computer interface dataset (BI2015b) (available at: https://hal.archives-ouvertes.fr/hal-02172347)
[31] Riccio A, Simione L, Schettini F, Pizzimenti A, Inghilleri M, Belardinelli M O, Mattia D and Cincotti F 2013 Attention and P300-based BCI performance in people with amyotrophic lateral sclerosis *Front. Hum. Neurosci.* **7** 732
[32] Farwell L A and Donchin E 1988 Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials *Electroencephalogr. Clin. Neurophysiol.* **70** 510–23
[33] Aricó P *et al* 2013 Influence of P300 latency jitter on event related potential-based brain-computer interface performance *J. Neural Eng.* **11** 035008
[34] Sosulski J and Tangermann M 2019 Electroencephalogram signals recorded from 13 healthy subjects during an auditory oddball paradigm under different stimulus onset asynchrony conditions *Dataset*
[35] Sosulski J and Tangermann M 2019 Spatial filters for auditory evoked potentials transfer between different experimental conditions *Graz BCI Conf.*
[36] Sosulski J *et al* 2021 Online optimization of stimulation speed in an auditory brain-computer interface under time constraints (arXiv:1706.06083)
[37] Ang K K, Chin Z Y, Wang C, Guan C and Zhang H 2012 Filter bank common spatial pattern algorithm on BCI competition IV datasets 2A and 2B *Front. Neurosci.* **6** 39

[38] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420
[39] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces *J. Neural Eng.* **15** 056013
[40] Sakhavi S, Guan C and Yan S 2018 Learning temporal information for brain-computer interface using convolutional neural networks *IEEE Trans. Neural Netw. Learn. Syst.* **29** 5619–29
[41] Mane R *et al* 2021FBCNet: a multi-view convolutional neural network for brain-computer interface (arXiv:2104.01233)
[42] Zhao H, Zheng Q, Ma K, Li H and Zheng Y 2021 Deep representation-based domain adaptation for nonstationary EEG classification *IEEE Trans. Neural Netw. Learn. Syst.* **32** 535–45
[43] Song Y, Zheng Q, Liu B and Gao X 2023 EEG conformer: convolutional transformer for eeg decoding and visualization *IEEE Trans. Neural Syst. Rehabil. Eng.* **31** 710–9
[44] Lin Z, Zhang C, Wu W and Gao X 2006 Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs *IEEE Trans. Biomed. Eng.* **53** 2610–4
[45] Salami A, Andreu-Perez J and Gillmeister H 2022 EEG-ITNet: an explainable inception temporal convolutional network for motor imagery classification *IEEE Access* **10** 36672–85
[46] Chen X *et al* 2023 Toward reliable signals decoding for electroencephalogram: a benchmark study to EEGNeX *Biomed. Signal Process. Control* **87** 105475
[47] Barachant A, Bonnet S, Congedo M and Jutten C 2012 Multiclass brain-computer interface classification by riemannian geometry *IEEE Trans. Biomed. Eng.* **59** 920–8
[48] Barachant A *et al* 2010 Riemannian geometry applied to BCI classification *Proc. 9th Int. Conf. Latent Variable Analysis and Signal Separation (LVA/ICA'10)* (Springer) pp 629–36
[49] Nakanishi M, Wang Y, Chen X, Wang Y-T, Gao X and Jung T-P 2018 Enhancing detection of SSVEPs for a high-speed brain speller using task-related component analysis *IEEE Trans. Biomed. Eng.* **65** 104–12
[50] Barachant A 2014 MEG decoding using riemannian geometry and unsupervised classification *Technical Report* (available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf)
[51] Chevallier S *et al* 2018 Brain-machine interface for mechanical ventilation using respiratory-related evoked potential *Proc. Artificial Neural Networks and Machine Learning - Icann 2018 (Lecture Notes in Computer Science)* vol 11141 V Kůrková, Y Manolopoulos, B Hammer, L Iliadis and I Maglogiannis (Springer)