POLITECNICO DI TORINO
Repository ISTITUZIONALE

Cross-Paradigm Consistency of EEG Decoding Pipelines: A Six-Dataset LOSO Benchmark of Handcrafted, Raw-EEG, and Topomap-Sequence Approaches

Original
Cross-Paradigm Consistency of EEG Decoding Pipelines: A Six-Dataset LOSO Benchmark of Handcrafted, Raw-EEG, and Topomap-Sequence Approaches / Ahmadi, H., Siahpoosh, M., Mesin, L.. - In: ALGORITHMS. - ISSN 1999-4893. - 19:9(2026). [10.3390/a19090806]

Availability:
This version is available at: 11583/3017900 since: 2026-11-13T17:23:48Z

Publisher:
MDPI

Published
DOI:10.3390/a19090806

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

21 November 2026

algorithms

MDPI

Article

Cross-Paradigm Consistency of EEG Decoding Pipelines: A Six-Dataset LOSO Benchmark of Handcrafted, Raw-EEG, and Topomap-Sequence Approaches

Hossein Ahmadi [1], Mahnaz Siahpoosh [2] and Luca Mesin [1,*]

[1] Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Turin, Italy; hossein.ahmadi@polito.it
[2] Department of Telecommunications, Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Jalal Al Ahmad Highway, Tehran 1411713116, Iran; m.siahpoosh@modares.ac.ir
* Correspondence: luca.mesin@polito.it

Abstract

Electroencephalography (EEG) decoding pipelines are commonly tailored to individual brain–computer interface (BCI) paradigms and datasets, making dataset-specific peak performance difficult to distinguish from consistency under a shared design. We examined whether one representation–architecture pipeline, reused without paradigm-specific architectural redesign but trained separately within each dataset, could remain competitive across motor imagery (MI), P300 event-related potential (ERP/P300), and steady-state visual evoked potential (SSVEP) tasks. Common handcrafted feature families, paradigm-specific representations and specialized classical methods, two established raw-EEG neural references (EEGNet and EEG Conformer), and a prespecified topomap-sequence implementation (TopoSeqNet) were evaluated on six public datasets using within-dataset leave-one-subject-out (LOSO) cross-validation. TopoSeqNet serves as the focal implementation of the shared topomap-sequence strategy, enabling the same representation-construction rule and architecture template to be evaluated across all three paradigms. Among the five principal comparison series, TopoSeqNet ranked first on two datasets and second on four. Its equal-dataset mean chance-corrected balanced accuracy was 0.682, compared with 0.654 for the dataset-selected specific/classical reference, 0.650 for EEG Conformer, 0.648 for EEGNet, and 0.438 for the dataset-selected common reference. Because three dataset entries share the OpenBMI acquisition framework, these summaries are descriptive; an acquisition-source-weighted sensitivity analysis yielded corresponding values of 0.645, 0.630, 0.617, 0.617, and 0.365. Together, these results demonstrate competitive cross-paradigm consistency of the complete TopoSeqNet pipeline under within-dataset subject generalization and establish a focused basis for future component-level, direct-transfer, interpretability, and deployment-efficiency studies.

Keywords: electroencephalography; brain–computer interface; motor imagery; P300; SSVEP; MOABB; cross-paradigm benchmarking; handcrafted features; topomap sequence

Check for updates

Academic Editors: Juvenal Rodriguez-Resendiz and Maryam Ravan

Received: 5 August 2026
Revised: 4 September 2026
Accepted: 18 September 2026
Published: 20 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Electroencephalography (EEG)-based brain–computer interfaces (BCIs) establish a direct mapping between patterns of neural activity and commands for communication, assistive control, rehabilitation, and human–machine interaction [1,2]. EEG is particularly attractive for non-invasive BCI development because it combines millisecond temporal

resolution with comparatively low cost, portability, and direct sensitivity to neural electrophysiology [1]. However, EEG decoding is complicated by low signal-to-noise ratios, non-stationarity, inter-subject and inter-session variability, and complex interactions across spatial, temporal, and spectral dimensions. Consequently, reliable decoding depends not only on the selected learning algorithm but also on how the EEG is represented before classification [2,3].

EEG representations are traditionally designed around the expected neural response of a particular BCI paradigm. Motor imagery (MI) pipelines commonly emphasize sensorimotor mu- and beta-band modulation and supervised spatial filtering through common spatial pattern (CSP) or filter-bank common spatial pattern (FBCSP) methods [4]. Systems based on the P300 event-related potential (ERP/P300) typically exploit stimulus-locked amplitudes, component latencies, and spatial filters such as xDAWN [5]. Steady-state visual evoked potential (SSVEP) recognition instead relies on power at the stimulation frequencies and their harmonics, often using canonical correlation analysis (CCA) or filter-bank canonical correlation analysis (FBCCA) [6,7]. These methods incorporate valuable neurophysiological prior knowledge and can perform extremely well when their assumptions match the task. Nevertheless, they produce incompatible feature spaces, require different preprocessing choices, and are not directly portable from one paradigm to another.

A related difficulty is that performance reported on a single dataset may not reflect the behavior of the same pipeline elsewhere. EEG datasets differ in participant populations, channel density, signal quality, class structure, trial duration, acquisition hardware, and experimental protocol. Consequently, a feature or classifier can appear highly effective under one configuration while performing poorly when the dataset changes. The Mother of All BCI Benchmarks (MOABB; version 1.5.0) was developed partly to address this problem by standardizing data access, preprocessing interfaces, evaluation procedures, and statistical comparisons [8,9]. Its benchmark analyses have shown that identical pipelines can produce markedly different outcomes across datasets and that a method validated on one dataset does not necessarily retain its advantage on another [8]. Thus, the field faces not a shortage of candidate algorithms, but uncertainty regarding which representation–model combinations remain dependable beyond the conditions for which they were initially designed.

Research has addressed this dependence from several directions. Within task-specific settings, ensemble learning has improved MI decoding by combining classifiers with complementary error patterns [10]. End-to-end convolutional neural networks (CNNs) learn temporal and spatial EEG features directly from the signal, reducing reliance on manual feature engineering [11]. EEGNet was explicitly designed as a compact deep-learning architecture applicable across several BCI paradigms [12], whereas EEG Conformer combines convolutional extraction with self-attention to capture local and longer-range dependencies [13]. Transfer-learning approaches based on Riemannian or Euclidean alignment further address inter-subject and inter-session variability by transforming participants' signal distributions into more comparable spaces [14,15].

More recent work has moved toward reusable EEG representations. Large pretrained models such as the Large Brain Model (LaBraM) learn generic representations from heterogeneous EEG datasets before downstream adaptation [16]. Task-independent encoders integrating CNNs, autoencoders, and Transformers have also been evaluated across MI, ERP/P300, and SSVEP data [3], while topology-preserving latent representations have revealed shared and condition-specific structure across multimodal imagination and perception tasks [2]. Related semantic-aware work on covert inner speech has used overt multimodal EEG–electromyography (EMG)–audio supervision, contrastive learning, and semantic-prototype alignment under a subject-held-out covert evaluation, illustrating another route to transferable EEG representations in a language-decoding setting [17]. Collectively, these

studies support the feasibility of learning broadly applicable EEG representations. The present study complements their emphasis on representation learning, pretraining, domain adaptation, robustness, latent-space interpretation, and overt–covert transfer by providing a controlled factorial assessment of whether conventional feature families and classifiers retain stable relative performance across physiologically distinct BCI paradigms and by distinguishing dataset-specific peak performance from cross-paradigm consistency.

Building on this literature, the present study treats cross-paradigm consistency as a distinct scientific criterion: can one representation-construction rule and architecture template, trained separately in each dataset under a common participant-held-out protocol, maintain competitive performance across MI, ERP/P300, and SSVEP? The factorial benchmark evaluates this criterion against common handcrafted representations, task-informed classical pipelines, and established raw-EEG neural references, thereby separating the stability of a shared design from dataset-specific peak accuracy.

Here, shared cross-paradigm design denotes reuse of the same representation rule and architecture template for MI, ERP/P300, and SSVEP. Paradigm-appropriate filtering and analysis intervals preserve each task's established signal requirements, and every model is trained and evaluated separately within its dataset. Accordingly, the benchmark quantifies generalization to unseen participants within each dataset; direct cross-dataset transfer provides the natural next level of evaluation.

Six publicly available datasets, with two datasets per paradigm, were evaluated through MOABB. Common handcrafted feature families were combined factorially with four conventional classifiers and compared with paradigm-specific representations and specialized classical methods. EEGNet and EEG Conformer provide complementary neural reference points: a compact cross-paradigm convolutional design and a convolution–attention architecture.

Topographic EEG sequences provide the focal shared representation. Each epoch is converted into an ordered sequence of scalp-voltage maps and processed by TopoSeqNet, a CNN–Transformer model that operationalizes this strategy. The article's scientific contribution is the controlled evaluation of this integrated representation–architecture pipeline across three physiologically distinct BCI paradigms. Evaluating the representation-construction rule and architecture jointly characterizes the behavior of the complete pipeline and motivates a focused component-level analysis of their interaction.

The main contributions of this study are as follows:

1. Common handcrafted features and four conventional classifiers are evaluated factorially across six MI, ERP/P300, and SSVEP dataset entries under the same participant-held-out framework, providing a controlled reference for dataset dependence.
2. Paradigm-specific representations and specialized classical methods are evaluated alongside two established raw-EEG neural references and one prespecified topomap-sequence pipeline, allowing dataset-specific peak performance to be distinguished from consistency of a shared design.
3. Dataset-level performance is complemented by participant-level paired comparisons, chance-corrected summaries, ranks among the five principal comparison series, and acquisition-source sensitivity analyses that assess the influence of the shared OpenBMI origin of three dataset entries.
4. Across the six dataset entries, the same TopoSeqNet pipeline remains within ranks 1–2, and its consistency is examined using equal-dataset, acquisition-source-weighted, and leave-one-source-out summaries.

The remainder of the paper is organized as follows. Section 2 describes the datasets, feature representations, classifiers, preprocessing, the focal topomap-sequence pipeline, the two raw-EEG neural references, and the evaluation protocol. Section 3 presents the con-

ventional reference results, principal pipeline comparisons, cross-dataset and acquisition-source summaries, and participant-level statistics. Section 4 interprets the cross-paradigm findings, their practical implications, and the next analyses motivated by the benchmark, and Section 5 summarizes the principal conclusions.

2. Methodology

2.1. Datasets

Six publicly available MOABB datasets covered MI, ERP/P300, and SSVEP, with two datasets per paradigm to reduce reliance on single-dataset conclusions [8,9]. The three OpenBMI datasets (Lee2019_MI, Lee2019_ERP, and Lee2019_SSVEP [18]) serve as a controlled multi-paradigm anchor under a largely common acquisition setting; the source-weighted analysis treats them as one acquisition source. Cho2017, BI2014a, and Wang2016 add independent variation in montage, hardware, sampling rate, protocol, and class structure [19–21]. Table 1 summarizes the datasets and the retained epochs after the common quality-control rule. Only the labeled offline training runs of Lee2019_MI and Lee2019_SSVEP were used because their MOABB online test runs lack trial labels.

Table 1. Characteristics of the EEG datasets and the intervals analyzed in this study. The recorded task/trial window is distinguished from the interval supplied to the evaluated pipelines. LH: left hand; RH: right hand; T: target; NT: non-target.

Panel A: Dataset characteristics

Paradigm	Dataset	Subjects	Sessions	Channels	f_s (Hz)	Classes
MI	Lee2019_MI [18]	54	2	62	1000	2 (LH, RH)
MI	Cho2017 [19]	52	1	64	512	2 (LH, RH)
ERP/P300	Lee2019_ERP [18]	54	2	62	1000	2 (T, NT)
ERP/P300	BI2014a [20]	64 [a]	1 [b]	16	512	2 (T, NT)
SSVEP	Lee2019_SSVEP [18]	54	2	62	1000	4 frequencies: 5.45, 6.67, 8.57, and 12 Hz
SSVEP	Wang2016 [21]	34 [c]	1	64	250	40 frequencies: 8–15.8 Hz

Panel B: Timing and retained epochs

Dataset	Recorded task/trial window	Analyzed interval	Retained epochs *n* (%)
Lee2019_MI	4 s	0–3 s after cue onset	10,638 (98.5%)
Cho2017	3 s	0–3 s after cue onset	10,236 (98.4%)
Lee2019_ERP	1 s (−0.2 to 0.8 s)	0–0.8 s after stimulus onset	445,330 (99.6%)
BI2014a	1 s (−0.2 to 0.8 s)	0–0.8 s after stimulus onset	73,673 (96.9%)
Lee2019_SSVEP	4 s	0–4 s after stimulation onset	10,728 (99.3%)
Wang2016	5 s of stimulation [d]	0–4 s for common features; 0–5 s for specific/classical and neural pipelines	8094 (99.2%)

[a] MOABB exposes 64 of the 71 participants reported for BI2014a; only the available participants were used. [b] The employed MOABB version represents BI2014a as a one-session dataset, although the original description reports up to three game sessions. [c] MOABB exposes 34 of the 35 participants reported for Wang2016; only the available participants were used. [d] Wang2016 stores 6-s epochs comprising 0.5 s before stimulation, 5 s of visual stimulation, and 0.5 s after stimulus offset. Note: Retention reflects the specified non-finite-value/500 μV peak-to-peak quality-control rule and is not an indicator of relative signal quality across datasets.

2.2. Paradigm-Agnostic Handcrafted Features

Predefined, formula-based temporal, spectral, complexity, and hemispheric features were calculated identically across all six datasets; common denotes identical definitions, not physiological neutrality [22–25]. Each family was evaluated separately and jointly as `Common_all`; dimensionality depended only on the available channels and homologous electrode pairs. Table 2 gives the exact implementation. All data-dependent preprocessing parameters were estimated exclusively from the training partition and applied unchanged to held-out data. All common features were calculated independently for each epoch without using class labels.

Table 2. Paradigm-agnostic handcrafted features used for all MI, ERP/P300, and SSVEP datasets. C and N denote the numbers of EEG channels and samples per epoch, respectively, and P denotes the number of available homologous electrode pairs.

Feature Family	Exact Implementation	Dimension
Time statistics	Mean, variance, standard deviation, root-mean-square amplitude, peak-to-peak amplitude, skewness, and excess kurtosis, computed independently for every channel and epoch. Skewness and kurtosis were calculated using bias-corrected estimates.	$7C$
Hjorth parameters [23]	Activity, mobility, and complexity computed from the signal and its first and second temporal differences.	$3C$
Absolute bandpower	Absolute power in δ (1–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz), and low-γ (30–40 Hz), obtained by integrating the Welch PSD over each band.	$5C$
Log-bandpower	Natural logarithm of absolute power, $\log(P_b + \varepsilon)$, in δ (1–4 Hz), θ (4–8 Hz), α (8–13 Hz), β (13–30 Hz), and low-γ (30–40 Hz).	$5C$
Relative bandpower	Power in each band divided by the total power between 1 and 40 Hz, $P_b/(P_{1\text{–}40} + \varepsilon)$.	$5C$
Spectral entropy [24]	Normalized Shannon entropy of the Welch PSD between 1 and 40 Hz.	C
Differential entropy [25]	Gaussian differential entropy estimated separately from the power in each of the five frequency bands as $\frac{1}{2}\log(2\pi e P_b + \varepsilon)$.	$5C$
Hemispheric asymmetry	Left-minus-right log-bandpower for every frequency band. Predefined pairs were: Fp1/Fp2, AF3/AF4, F7/F8, F5/F6, F3/F4, F1/F2, FC5/FC6, FC3/FC4, FC1/FC2, C5/C6, C3/C4, C1/C2, CP5/CP6, CP3/CP4, CP1/CP2, P7/P8, P5/P6, P3/P4, P1/P2, PO7/PO8, PO3/PO4, and O1/O2. Only pairs available in a given montage were retained.	$5P$
`Common_all`	Concatenation of all feature families above.	$31C + 5P$

Throughout the feature calculations, the numerical constant $\varepsilon = 10^{-12}$ was added to logarithm arguments and denominators to avoid undefined numerical values caused by zero power, variance, or probability mass. Welch power spectral density (PSD) estimates [22] used a Hann window, segment length $\min(N, \text{round}(f_s))$ samples (1 s when permitted), 50% overlap, and density scaling; PSD samples were not used directly as classifier inputs.

2.3. Paradigm-Specific Features and Classical Baselines

In addition to the common representation, features based on the characteristic neural response of each paradigm were evaluated. MI features emphasized sensorimotor mu- and beta-band modulation, asymmetry, and within-epoch event-related desynchronization/synchronization (ERD/ERS)-related changes; ERP/P300 features preserved stimulus-locked waveform amplitudes, latencies, and xDAWN-based spatial information; and SSVEP features measured fundamental and harmonic power, narrowband signal-to-noise ratio (SNR), phase-locking value (PLV), and similarity to sinusoidal frequency references. These

representations provide strong task-informed baselines against which the consistency of the paradigm-agnostic and topomap-based approaches can be assessed.

Deterministic feature blocks were evaluated individually and as a paradigm-specific concatenated representation. CSP/FBCSP, xDAWN-based, CCA, and FBCCA methods were additionally included as specialized classical methods. Their exact configurations are reported in Table 3.

Table 3. Paradigm-specific representations and classical reference methods evaluated in the study.

Task	Method	Exact Implementation
MI	Mu-band power; Beta-band power; Mu+beta bandpower; ERD/ERS; Sensorimotor asymmetry; `MI_specific_all`	Mu- and beta-band representations contained channel-wise log-power and relative power at 8–13 Hz and 13–30 Hz, with relative power normalized by the combined mu-plus-beta power. The Mu+beta bandpower block concatenated the mu- and beta-band representations. The ERD/ERS proxy was late-minus-early log-power, comparing 0.5–3.0 s with 0–0.5 s; it therefore represents a within-epoch temporal contrast rather than a prestimulus-baseline measure. Sensorimotor asymmetry was left-minus-right log-power over the predefined motor-region pairs. `MI_specific_all` concatenated Mu+beta bandpower, asymmetry, and ERD/ERS without duplicating the separate single-band blocks.
MI	CSP; FBCSP [4]	CSP retained four components from an 8–30 Hz copy. FBCSP retained four components from each of five bands: 8–12, 12–16, 16–20, 20–24, and 24–30 Hz. Both used fourth-order zero-phase Butterworth filtering, Ledoit–Wolf covariance regularization, log-average-power transformation, and no trace normalization.
ERP	ERP window means; P200 amplitude; N200 amplitude; P300 amplitude; P300 latency; `ERP_specific_all`	ERP window means were calculated in the 0–100, 100–180, 180–280, 280–450, 450–650, and 650–800 ms intervals. P200 and N200 amplitudes were the positive and negative channel-wise peaks at 150–250 and 150–300 ms, respectively. P300 amplitude and latency were calculated in the 250–600 and 300–700 ms intervals, respectively, with additional centro-parietal mean and maximum amplitude, peak latency, and spatial standard deviation. Amplitude- and latency-derived quantities were evaluated separately. `ERP_specific_all` concatenated all preceding ERP blocks.
ERP	xDAWN covariance; xDAWN [5]	Four supervised xDAWN filters were fitted within each training fold. xDAWN covariance used covariance features from the filtered epochs, whereas xDAWN used ERP-window and P300 amplitude/latency descriptors.
SSVEP	Fundamental power; Harmonic power; Fundamental + harmonics; Narrowband SNR; PLV; `SSVEP_specific_all`	Posterior-channel log-power was calculated from a Hann-windowed full-epoch periodogram at the frequency bin nearest each stimulation frequency and each valid second and third harmonic not exceeding 45 Hz. Fundamental power, Harmonic power, and Fundamental + harmonics used the fundamental, harmonics, or their concatenation. Narrowband SNR was the natural logarithm of the ratio between power at the nearest target or harmonic bin and the mean power across the eight flanking bins at offsets ± 2 through ± 5, excluding the immediately adjacent bins. PLV measured phase consistency between posterior EEG activity and the corresponding sinusoidal reference; channel-wise and posterior summary values were retained. `SSVEP_specific_all` concatenated Fundamental + harmonics, Narrowband SNR, and PLV.
SSVEP	CCA features; CCA [6]; FBCCA features; FBCCA [7]	CCA used sine and cosine references at f, $2f$, and $3f$. CCA features denote the one-score-per-class vectors supplied to the conventional classifiers, whereas CCA denotes direct maximum-score recognition. FBCCA used three fourth-order zero-phase Butterworth filter banks: 4–45, 8–45, and 14–45 Hz. Squared CCA correlations were combined using $w_m = m^{-1.25} + 0.25$, where m is the filter-bank index, with weights normalized to sum to one. FBCCA features denote the class-score vectors supplied to the classifiers, whereas FBCCA denotes direct maximum-score recognition.

Channel sets: Motor-region channels: FC1–FC6, FCz, C1–C6, Cz, CP1–CP6, and CPz. Motor-region pairs: FC1/FC2, FC3/FC4, FC5/FC6, C1/C2, C3/C4, C5/C6, CP1/CP2, CP3/CP4, and CP5/CP6. Centro-parietal channels: FC1, FC2, FCz, C1–C4, Cz, CP1–CP4, CPz, P1–P4, Pz, PO3, PO4, POz, and Oz. Posterior channels: O1, O2, Oz, PO3–PO8, POz, P1–P8, and Pz. Only available channels or pairs were retained. All supervised spatial transformations were fitted exclusively within the corresponding training fold.

The suffix `_specific_all` denotes the concatenation of the preceding deterministic feature blocks for the corresponding paradigm. All representations except direct CCA and direct FBCCA were evaluated with the four conventional classifiers. `CCA features` and `FBCCA features` denote the corresponding score vectors used as classifier inputs, whereas CCA and FBCCA denote native maximum-score recognition.

2.4. Machine-Learning Classifiers

Four conventional classifiers were evaluated for every tabular representation: shrinkage linear discriminant analysis (sLDA), linear support vector machine (Linear SVM), SVM with a radial basis function kernel (RBF-SVM), and extreme gradient boosting (XGBoost), representing linear generative, linear discriminative, nonlinear kernel, and tree-ensemble models, respectively [26–29]. Hyperparameters were fixed a priori across datasets and feature families, except for the binary or multiclass XGBoost objective and the ERP/P300 class-imbalance adjustment. Table 4 lists the configurations.

Table 4. Fixed configurations of the conventional classifiers. K denotes the number of classes.

Classifier	Configuration
sLDA	`solver=lsqr`; `shrinkage=auto`; equal class priors, $\pi_k = 1/K$
Linear SVM	`kernel=linear`; $C = 1$; `class_weight=balanced`; probability estimates enabled
RBF-SVM	`kernel=rbf`; $C = 3$; $\gamma =$ `scale`; `class_weight=balanced`; probability estimates enabled
XGBoost	300 trees; maximum depth = 3; learning rate = 0.03; row subsampling = 0.9; feature subsampling = 0.9; minimum child weight = 1; $\gamma = 0$; L_1 regularization = 0; L_2 regularization = 1; binary logistic objective for MI and ERP/P300; multiclass soft-probability objective for SSVEP; ERP positive-class weight = N_{NT}/N_T, where N_{NT} and N_T denote the numbers of non-target and target training trials, respectively

Missing feature values were replaced by the median estimated from the corresponding training partition, and zero-variance features were removed. Features supplied to sLDA and the two SVM classifiers were standardized using the training-fold mean and standard deviation. The fitted preprocessing transformations were subsequently applied unchanged to the held-out data. Class-balanced weights were used for both SVM classifiers. For ERP/P300, the XGBoost positive-class weight was calculated within each training fold as the ratio of non-target to target trials. No preprocessing parameter or class statistic was estimated from validation or test data.

The four classifiers were applied to every non-direct tabular representation, including CSP, FBCSP, xDAWN covariance, xDAWN, CCA features, and FBCCA features. Direct CCA and direct FBCCA were evaluated separately using their native maximum-score recognition rules. The topomap-based and raw-EEG deep-learning models were likewise evaluated separately as end-to-end representation-learning approaches.

2.5. Preprocessing

Having defined the evaluated feature representations and classifiers, this subsection describes the common and paradigm-dependent signal-processing procedures used to construct their inputs. All datasets were accessed through MOABB version 1.5.0 [8,9]. The MOABB loaders convert the stored signals to volts, construct recording objects using MNE-Python version 1.12.1, and attach the available channel, montage, and event information. The three Lee2019 datasets and BI2014a are exposed without dataset-level digital filtering. The Cho2017 loader temporally demeans the originally segmented trials before reconstructing a continuous recording, whereas Wang2016 is distributed as pre-epoched data that were already downsampled from 1000 to 250 Hz; its MOABB loader similarly

demeans the individual trials. These loader-level operations were retained and were not repeated in the subsequent pipeline.

The additional preprocessing was designed to reduce acquisition-related variability while preserving the paradigm-specific neural responses required by the evaluated feature families and classical baselines. All datasets were therefore resampled to a common rate of 250 Hz and re-referenced to the common average. Filtering ranges and analysis intervals were then adapted to the temporal and spectral characteristics of MI, ERP/P300, and SSVEP, as reported in Table 5. Separate signal copies were generated for CSP/FBCSP, ERP-specific and xDAWN-based representations, and SSVEP power, SNR, PLV, CCA, and FBCCA representations, as summarized in Table 5.

Table 5. Preprocessing applied after loading the datasets through MOABB. Broad filtering was performed through the MOABB/MNE paradigm pipeline, whereas additional paradigm-specific signal copies were generated using fourth-order zero-phase Butterworth filters.

Stage	**Motor Imagery**	**ERP/P300**	**SSVEP**
Datasets	Lee2019_MI, Cho2017	Lee2019_ERP, BI2014a	Lee2019_SSVEP, Wang2016
Broad band-pass	1–40 Hz	0.5–40 Hz	1–45 Hz
Epoch interval	0–3 s relative to MI cue onset	−0.2–0.8 s relative to stimulus onset; all analyses restricted to 0–0.8 s	0–4 s for common features; complete stimulation interval for SSVEP-specific and deep-learning methods (4 s for Lee2019_SSVEP and 5 s for Wang2016)
Baseline correction	None	Channel-wise subtraction of the mean over the −0.2–0 s interval	None
Paradigm-specific copy	8–30 Hz for CSP; five FBCSP bands: 8–12, 12–16, 16–20, 20–24, and 24–30 Hz	0.5–20 Hz for ERP-specific, xDAWN covariance, and xDAWN representations	4–45 Hz for fundamental/harmonic power, narrowband SNR, PLV, CCA, and FBCCA
Operations common to all paradigms			
Sampling rate	250 Hz		
Reference	Common-average reference applied after epoch extraction		
Artifact rejection	Rejection of epochs containing non-finite values or any channel with peak-to-peak amplitude greater than 500 µV		
Training-fold operations	Median imputation; zero-variance feature removal; `SelectKBest` using the ANOVA F-statistic (`f_classif`) with $k = \min(1000, D)$ when $D > 1000$, where D is the input feature dimension; and feature scaling for sLDA and SVM classifiers		

A common artifact-rejection criterion was applied after epoch extraction and re-referencing. Epochs containing non-finite values or a peak-to-peak amplitude greater than 500 µV in any EEG channel were discarded. Independent component analysis (ICA) and electrooculography (EOG) regression procedures were not used because dedicated ocular channels are not consistently available across the six datasets; applying such correction only to selected datasets would introduce an additional source of preprocessing heterogeneity. The numbers of accepted and rejected epochs were recorded for every dataset. Accordingly, the scalp-voltage maps represent sensor-level activity; residual ocular or other non-task activity may remain, and interpretation is confined to the sensor level rather than to direct maps of cortical activation.

All operations that estimated parameters from the observed data distribution were confined to the training partition of each evaluation fold. These included missing-value imputation, removal of zero-variance features, feature selection, feature scaling, and the fitting of supervised spatial transformations such as FBCSP and xDAWN. The fitted transformations were subsequently applied unchanged to the held-out data.

2.6. Shared Topomap-Sequence Pipeline and Raw-EEG Neural References

The topomap-sequence pipeline operationalizes a shared cross-paradigm design. TopoSeqNet is the CNN–Transformer used to process the ordered map sequences. Reusing one representation rule and architecture template across MI, ERP/P300, and SSVEP allows the study to test whether the integrated pipeline maintains competitive LOSO performance across heterogeneous signal structures. Accordingly, the primary unit of interpretation is the complete representation–architecture pipeline, with component-level attribution reserved for a dedicated controlled study.

Each analyzed EEG epoch was converted into a temporally ordered sequence of scalp-voltage maps. At 250 Hz, non-overlapping blocks of five samples were averaged, yielding one map every 20 ms, with frame centers at $8 + 20k$ ms, where $k = 0, \ldots, T - 1$, and T is the number of maps in the epoch. Electrode locations obtained through MOABB/MNE were completed using the `standard_1005` montage when necessary, normalized to the unit sphere, projected onto a two-dimensional scalp plane, centered, and scaled to a maximum sensor radius of 0.92. Delaunay barycentric interpolation mapped channel values to grid points within the electrode convex hull; points inside the scalp mask but outside the convex hull were assigned the value measured at the nearest electrode, and points outside the scalp mask were set to zero. The resulting sequence contained 150 maps for each MI epoch, 40 for each ERP/P300 epoch, 200 for each Lee2019_SSVEP epoch, and 250 for each Wang2016 epoch.

The 32×32 grid and five-sample temporal aggregation were prespecified engineering choices and were held fixed across all six dataset entries. The grid provides a common, moderately sized spatial array for montages with different channel counts, while five-sample averaging reduces the sequence length and sample-scale variation while retaining an ordered 20 ms description of the epoch. These fixed settings make the benchmark a direct test of one prespecified construction rule across montage densities. Interpolation estimates values between measured electrodes, so its spatial resolution reflects the density and coverage of the recorded montage. A focused study across grid resolutions, temporal aggregation windows, and interpolation rules can extend this benchmark by quantifying design stability.

For each outer fold, one global scalar mean and standard deviation were estimated across all samples and channels of the fitting-participant epochs and used to standardize the inputs to all three neural pipelines before model-specific representation construction. The same values were applied unchanged to the validation and held-out test epochs. The resulting maps were supplied to TopoSeqNet as single-channel floating-point voltage arrays rather than RGB images. TopoSeqNet applies a shared two-dimensional CNN with squeeze-and-excitation blocks to each map, followed by global average pooling and a lightweight Transformer operating on the ordered map embeddings. One class-probability vector is returned for the complete epoch. Figure 1 traces the complete sequential data path evaluated in the benchmark.

EEGNet [12] and EEG Conformer [13] were selected as complementary raw-EEG reference architectures, spanning a compact convolutional design and a convolution–attention design. All three neural pipelines received the same broadly preprocessed epochs, returned one prediction per epoch, and used study-wide configurations across dataset entries

apart from dataset-forced input dimensions, sequence length, and class count. The same fitting/validation participant split, random seeds, loss, optimizer schedule, and stopping rule were applied within each outer fold. This common protocol supports a controlled comparison of predictive consistency while preserving each architecture's native input organization.

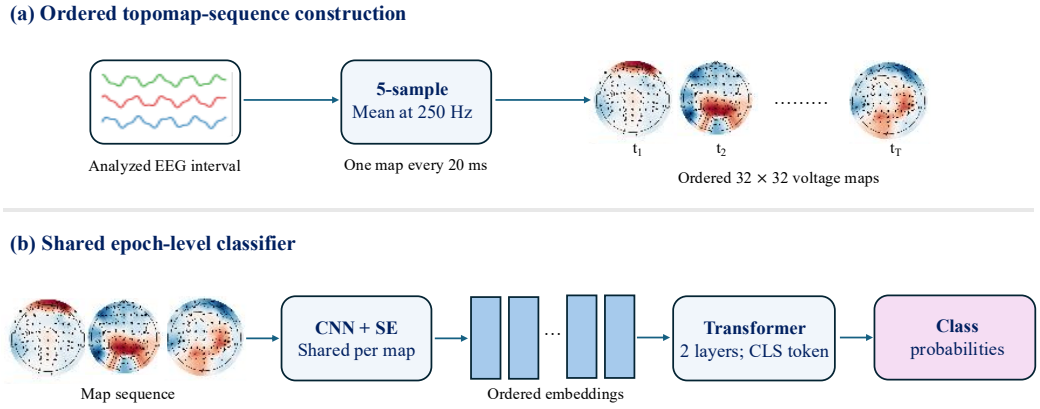


Figure 1. Construction and processing of one topomap-sequence epoch. After dataset- and paradigm-specific preprocessing, non-overlapping five-sample blocks are averaged at 250 Hz; electrode locations are projected onto the scalp plane, and the averaged channel values are interpolated using the procedure described in Section 2.6 to form an ordered sequence of single-channel 32 × 32 scalp-voltage arrays at 20-ms intervals. A shared CNN/squeeze-and-excitation encoder processes every map, and a Transformer integrates the ordered embeddings to produce one epoch-level class-probability vector. Sequence lengths are 150 for MI, 40 for ERP/P300, 200 for Lee2019_SSVEP, and 250 for Wang2016.

The TopoSeqNet configuration was obtained by adapting the published framework in [30] to epoch-level sequence classification and was developed independently of all six benchmark entries. The configuration was fixed before the reported outer-fold evaluations and applied uniformly across all six entries, without dataset-specific architecture or hyperparameter selection. EEGNet and EEG Conformer were instantiated from their cited Braindecode 1.6.1 implementations using the fixed configurations reported in Table 6.

Paradigm-appropriate filtering, analysis intervals, sequence lengths, and output dimensions follow each dataset's task structure, while training and evaluation use the shared protocol described below. Every model is fitted separately within a dataset, so the benchmark targets generalization to unseen participants within that dataset; direct cross-dataset transfer is an important extension.

The topomap-sequence representation builds on our previous work [30]; the present study extends that foundation from single-dataset frame classification to epoch-level cross-paradigm consistency. In [30], each perception or imagination epoch produced exactly 200 RGB 150 × 150 topographic images, generated at 15 ms intervals for perception and 20 ms intervals for imagination, and a CNN with squeeze-and-excitation blocks classified each frame independently. Here, each analyzed epoch is represented by 40–250 single-channel 32 × 32 voltage maps generated every 20 ms; a shared CNN encoder with squeeze-and-excitation blocks maps each frame to an embedding, and a two-layer Transformer integrates the ordered embeddings to produce one prediction for the complete epoch. The earlier model was trained with Adam using a learning rate of 10^{-4}, a batch size of 32, L_2 regularization, and up to 50 epochs. The present pipeline uses AdamW with a learning rate of 10^{-3}, weight decay of 10^{-4}, an effective batch size of 64, and up to 200 epochs, and averages class probabilities from three independently seeded models. Scientifically, ref. [30] examined subject-independent pictorial perception-versus-imagination decoding in

one 15-participant dataset, whereas the present study evaluates the cross-paradigm consistency of one prespecified epoch-level pipeline across six public MI, ERP/P300, and SSVEP benchmark entries using separate within-dataset LOSO evaluations. The scientific unit of evaluation is therefore the complete representation–architecture pipeline and its consistency across heterogeneous BCI paradigms.

Table 6. Prespecified configurations of the focal TopoSeqNet pipeline and the two raw-EEG neural reference models. Each model receives one complete analyzed epoch and returns one epoch-level prediction. The configurations are fixed across datasets except for input-channel count, epoch/sequence length, and number of output classes.

Model	Input Representation	Architecture Configuration
TopoSeqNet	Sequence of single-channel 32×32 topomaps generated every 20 ms	Three shared 3×3 convolutional blocks with 16, 32, and 64 filters; batch normalization, Gaussian error linear unit (GELU), and squeeze-and-excitation with reduction ratio 4; 2×2 max pooling and dropout = 0.10 after the first two blocks; global average pooling; embedding dimension = 64; learnable classification token and positional embeddings; two Transformer layers; four attention heads; feed-forward dimension = 128; Transformer dropout = 0.20; classifier dropout = 0.30.
EEGNet [12]	Raw multichannel EEG epoch, $C \times N$	Braindecode `EEGNet`: $F_1 = 8$; depth multiplier $D = 2$; $F_2 = 16$; temporal kernel length = 64 samples; depthwise kernel length = 16; pooling kernels = 4 and 8; mean pooling; spatial-convolution maximum norm = 1; dropout = 0.25; automatically determined final convolution length.
EEG Conformer [13]	Raw multichannel EEG epoch, $C \times N$	Braindecode `EEGConformer`: 40 temporal filters; temporal kernel length = 25 samples; average-pooling length = 75 and stride = 15; six Transformer layers; ten attention heads; convolutional dropout = 0.50; attention dropout = 0.50; automatically determined final fully connected length.

2.7. Evaluation Protocol and Statistical Analysis

All handcrafted, classical, topomap-based, and raw-EEG deep-learning pipelines were evaluated using subject-independent LOSO cross-validation, applied separately to each dataset. In every outer fold, all epochs of one participant formed the test set, and no information from that participant was used for preprocessing estimation, adaptive feature extraction, classifier fitting, neural-network training, early stopping, or model selection.

Every common and paradigm-specific non-direct representation was evaluated with the four conventional classifiers. CSP, FBCSP, xDAWN covariance, and xDAWN were fitted exclusively on the outer training participants and applied unchanged to the held-out participant. CCA features and FBCCA features were supplied to the classifiers, whereas direct CCA and direct FBCCA selected the class associated with the maximum reference-matching score.

For each deep-learning model, the outer training participants were divided into fitting and validation subsets. The number of validation participants was fixed as

$$N_{\text{val}} = \lceil 0.15 N_{\text{train}} \rceil, \tag{1}$$

where N_{train} is the number of participants remaining after exclusion of the outer test participant. Validation participants were selected using a prespecified deterministic fold-specific seed schedule (2026 plus the one-based outer-fold number) and were identical for TopoSeqNet, EEGNet, and EEG Conformer within each outer fold. A nested inner LOSO procedure was not used because the model architectures and training hyperparameters were fixed a

priori; the validation subset was required only for early stopping and checkpoint selection. No data augmentation or pretrained weights were used.

Each neural model was trained using seeds 42, 43, and 44, and the resulting class probabilities were averaged for the final test prediction. Weighted cross-entropy was used, with class weights estimated from the fitting participants. AdamW was applied with a learning rate of 10^{-3} and a weight decay of 10^{-4}. An effective batch size of 64 was implemented using microbatch/gradient-accumulation pairs of 8/8 for TopoSeqNet, 64/1 for EEGNet, and 32/2 for EEG Conformer. Training was limited to 200 epochs, and gradients were clipped at a norm of 1.0. The learning rate was reduced by a factor of 0.5 after seven epochs without improvement in validation balanced accuracy, with a minimum value of 10^{-5}. Early stopping was triggered after 20 epochs without improvement, and the best validation checkpoint was restored.

Balanced accuracy was used as the single primary metric for every pipeline:

$$\mathrm{BA} = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \tag{2}$$

where K is the number of classes, and TP_k and FN_k denote the numbers of true-positive and false-negative predictions for class k, respectively. This metric assigns equal importance to all classes and is therefore applicable to both balanced and imbalanced tasks. One participant-level balanced-accuracy value was obtained from each outer fold.

Because the datasets contain two, four, or forty classes, chance-corrected balanced accuracy, denoted $\mathrm{BA}_{\mathrm{cc}}$, was additionally used for cross-dataset summaries:

$$\mathrm{BA}_{\mathrm{cc}} = \frac{\mathrm{BA} - 1/K}{1 - 1/K}. \tag{3}$$

Raw balanced accuracy remained the principal within-dataset measure; $\mathrm{BA}_{\mathrm{cc}}$ provided a descriptive normalization across chance levels, while statistical interpretation remained dataset specific. The equal-dataset mean weighted all six entries equally, regardless of sample size or uncertainty.

Statistical analyses were conducted on paired participant-level scores. Four participant-level comparisons were examined separately within each dataset: TopoSeqNet versus EEGNet, EEG Conformer, the dataset-selected common reference, and the dataset-selected specific/classical reference. The EEGNet and EEG Conformer comparisons were defined a priori. Each selected reference was chosen separately within its dataset as the evaluated pipeline with the highest observed mean participant-level balanced accuracy. Comparisons with these two data-selected targets are therefore exploratory and conditional on post hoc comparator selection; they do not have the same confirmatory status as the two predefined neural-model comparisons.

For each comparison, the mean paired balanced-accuracy difference and its 95% bias-corrected and accelerated (BCa) bootstrap confidence interval (CI) were estimated using 10,000 participant-level resamples. Statistical significance was assessed using a two-sided sign-flipping permutation test, using the exact distribution when feasible and 100,000 random permutations otherwise. Holm correction was applied within each dataset across the four headline comparisons, and rank-biserial correlation was reported as the paired effect size. The BCa confidence intervals were not multiplicity-adjusted; multiplicity control was applied to the permutation-test p-values. Cross-dataset consistency was summarized using chance-corrected balanced accuracy and average method rank; epochs were never pooled across participants or datasets for statistical inference.

Two post hoc descriptive sensitivity summaries were additionally calculated from the existing dataset-level results without retraining any pipeline. First, to quantify the leverage

of Wang2016 on the equal-dataset summary, chance-corrected balanced accuracy was averaged over the other five dataset entries after omitting Wang2016. Second, to reduce the influence of the three related OpenBMI entries, their three chance-corrected values were first averaged within each principal series, and this OpenBMI value was then weighted equally with Cho2017, BI2014a, and Wang2016. Leave-one-acquisition-source-out summaries were also calculated over these four sources. With four acquisition sources, the role of these sensitivity analyses is descriptive rather than inferential.

3. Results

Across the six datasets, 558,699 of 563,312 epochs (99.18%) were retained after preprocessing and artifact rejection. Dataset-specific retention ranged from 96.9% for BI2014a to 99.6% for Lee2019_ERP as reported in Table 1. These percentages document the application of the specified quality-control rule and must not be interpreted as comparative measures of dataset signal quality.

3.1. Conventional Pipelines as Benchmark References

Figure 2 establishes the common handcrafted reference. The dataset-selected combination was Common_all with RBF-SVM for Lee2019_MI (0.681), log-bandpower with RBF-SVM for Cho2017 (0.672), time statistics with XGBoost for Lee2019_ERP (0.816), Common_all with sLDA for BI2014a (0.681), log-bandpower with RBF-SVM for Lee2019_SSVEP (0.815), and Common_all with XGBoost for Wang2016 (0.194). The change in the selected feature–classifier combination across datasets provides the empirical reference for testing the consistency of shared designs.

Figure 3 provides the corresponding task-informed references. The dataset-selected specific/classical pipelines were FBCSP with sLDA for Lee2019_MI (0.717) and Cho2017 (0.698), xDAWN with Linear SVM for Lee2019_ERP (0.873), xDAWN with sLDA for BI2014a (0.762), and direct FBCCA for Lee2019_SSVEP (0.942) and Wang2016 (0.901). These task-informed results provide the complementary benchmark reference against which the shared representation–architecture strategy is evaluated.

The Wang2016 common-reference result should be interpreted in the context of its specified input configuration. The common-feature pipeline used a 4 s interval and approximately 1 Hz Welch resolution, whereas the specialized and neural pipelines used the full 5 s stimulation interval, and the SSVEP classes are spaced by approximately 0.2 Hz. Accordingly, this result quantifies the performance of the specified common-feature configuration, with signal duration and Welch resolution forming part of that configuration.

3.2. Comparison with Deep-Learning Models

Figure 4 and Table 7 compare the principal pipelines. TopoSeqNet ranked first on Lee2019_MI (0.741) and Lee2019_ERP (0.934), and second on the other four datasets. On Cho2017, its mean balanced accuracy was numerically close to that of FBCSP (0.697 versus 0.698); it trailed EEGNet on BI2014a and remained close to direct FBCCA on both SSVEP datasets while exceeding both raw-EEG neural references there. Its rank therefore never fell below second despite substantial differences in paradigm, montage, class count, and protocol.

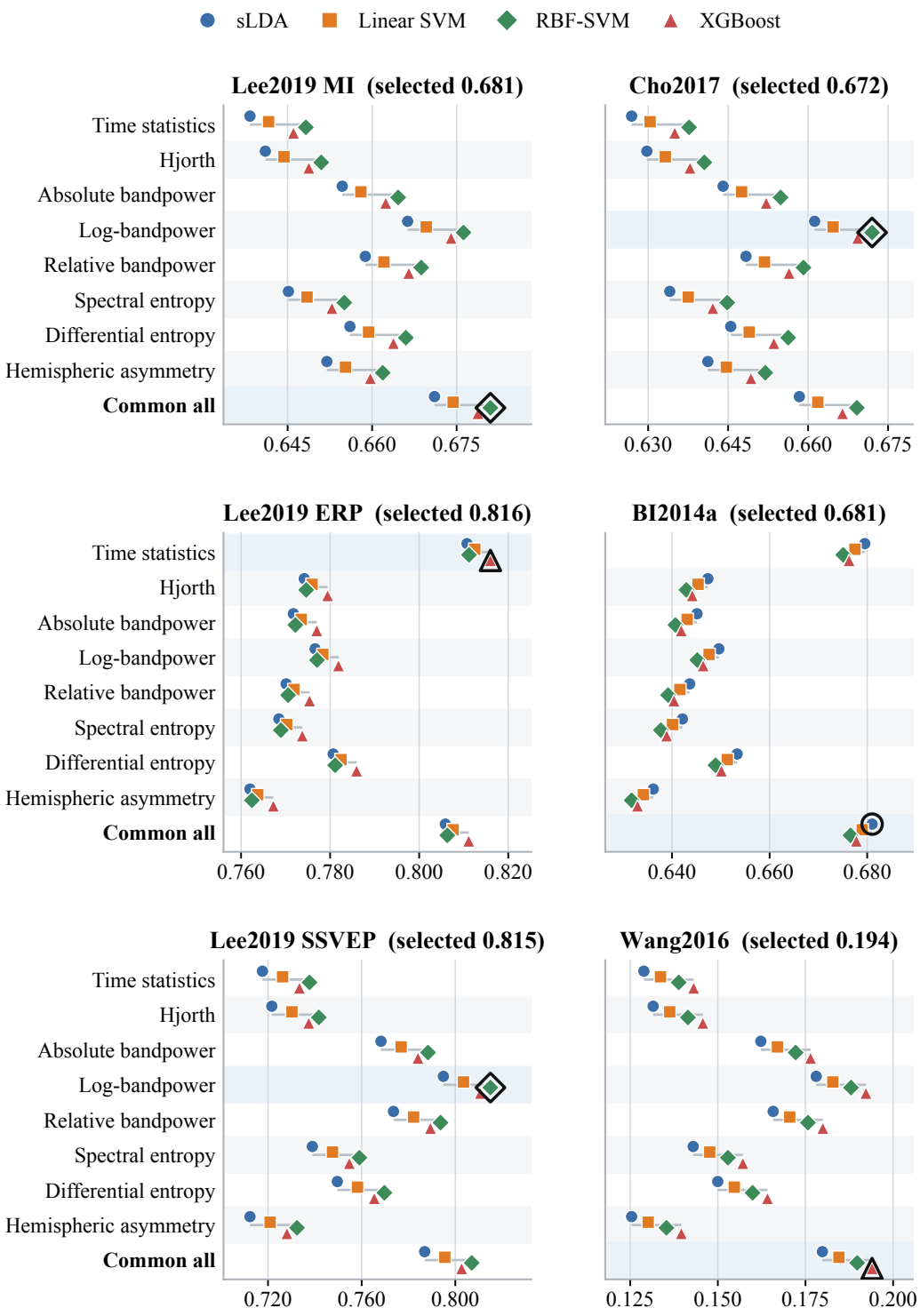


Figure 2. Mean within-dataset LOSO balanced accuracy for every common handcrafted feature-family and classifier combination. Markers identify classifiers, and horizontal segments span their results for each feature family. Horizontal scales are dataset-specific to resolve within-dataset differences and must not be compared geometrically across panels. The outlined marker identifies the dataset-selected common reference, defined as the highest observed mean among the 36 common feature–classifier combinations evaluated in that dataset.

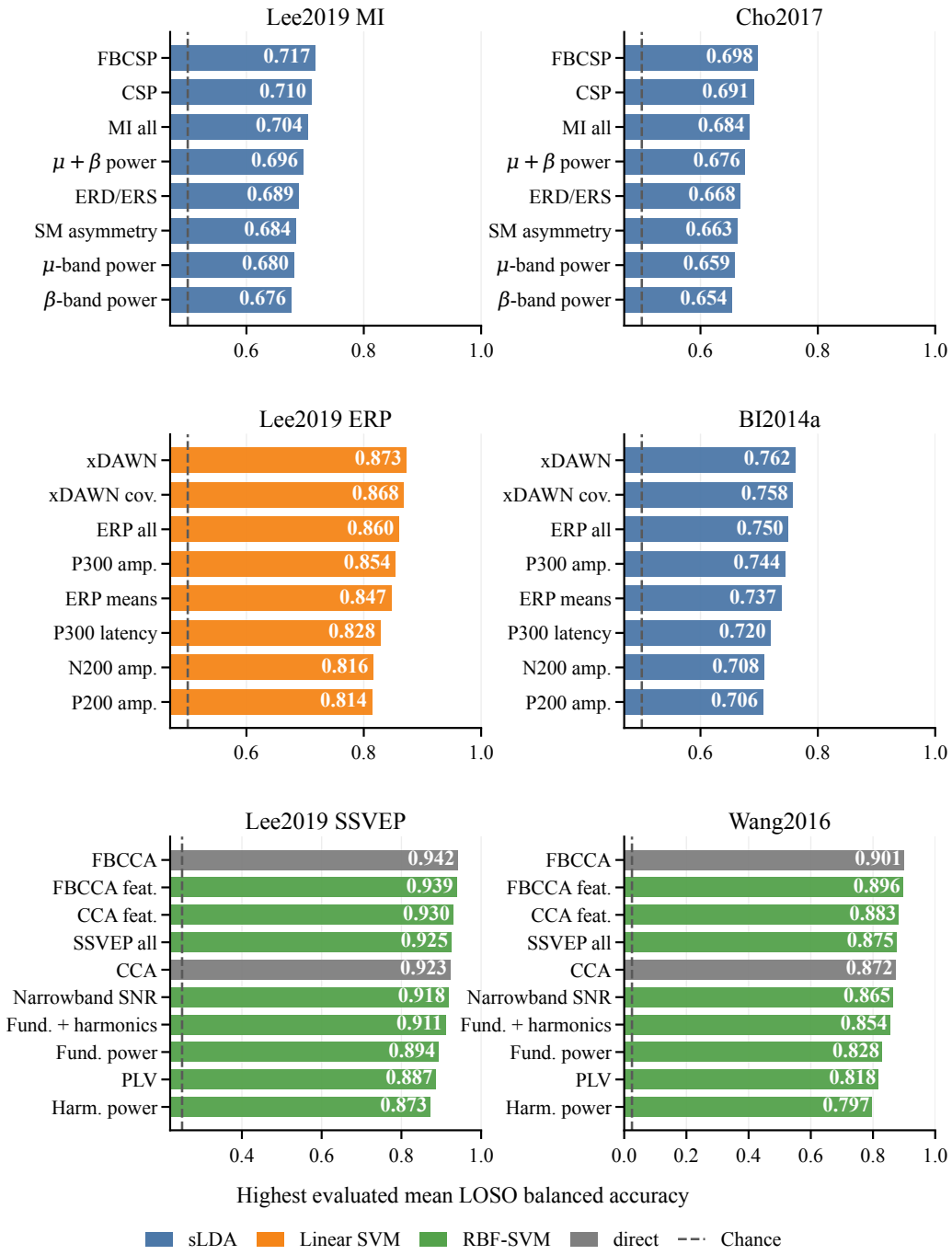


Figure 3. Performance of the evaluated paradigm-specific representations and specialized classical methods. For classifier-based representations, each bar shows the highest mean within-dataset LOSO balanced accuracy obtained across the four conventional classifiers, and bar identifies the selected classifier. Gray bars for CCA and FBCCA show native direct maximum-score recognition and are therefore not classifier-selected; their feature variants are classifier inputs. Dashed vertical lines indicate chance performance. These methods provide the study's predefined set of specialized benchmark references. Abbreviations: SM, sensorimotor; cov., covariance; amp., amplitude; feat., features; Fund., fundamental; Harm., harmonic.

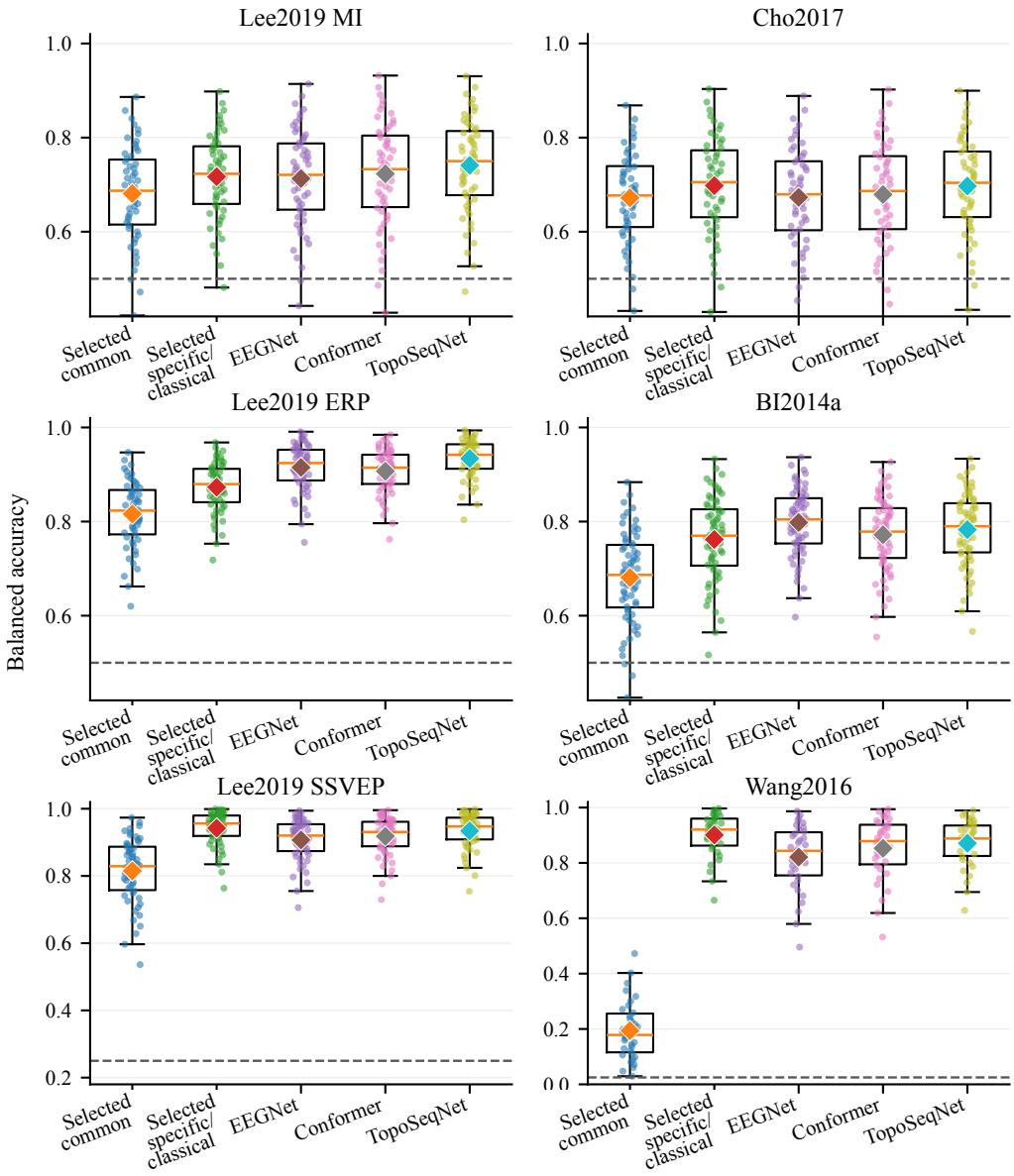


Figure 4. Participant -level balanced-accuracy distributions for the principal pipelines. Each point denotes one held-out participant's balanced-accuracy score, boxes the distributions, diamonds the means, and dashed lines chance-level accuracy.

Table 7. Participant-level balanced accuracy of the five principal comparison series. Values are mean ± standard deviation across held-out participants. Dataset-selected reference values report the highest observed mean among the evaluated pipelines within each category, whereas the neural-network columns correspond to fixed pipelines. Standard deviations summarize participant heterogeneity. Bold values indicate the highest mean within each dataset row.

	Dataset-Selected References		**Neural-Network Pipelines**		
Dataset	**Common**	**Specific/Classical**	**EEGNet**	**EEG Conformer**	**TopoSeqNet**
Lee2019_MI	0.681 ± 0.102	0.717 ± 0.091	0.713 ± 0.104	0.723 ± 0.112	**0.741 ± 0.101**
Cho2017	0.672 ± 0.096	**0.698 ± 0.105**	0.673 ± 0.108	0.679 ± 0.114	0.697 ± 0.103
Lee2019_ERP	0.816 ± 0.071	0.873 ± 0.054	0.915 ± 0.051	0.907 ± 0.048	**0.934 ± 0.041**
BI2014a	0.681 ± 0.098	0.762 ± 0.089	**0.798 ± 0.072**	0.772 ± 0.079	0.783 ± 0.078
Lee2019_SSVEP	0.815 ± 0.097	**0.942 ± 0.051**	0.907 ± 0.063	0.918 ± 0.058	0.934 ± 0.053
Wang2016	0.194 ± 0.106	**0.901 ± 0.079**	0.821 ± 0.119	0.853 ± 0.112	0.871 ± 0.086

3.3. Cross-Dataset Consistency

Because chance-level balanced accuracy differs between the binary, four-class, and 40-class datasets, cross-dataset consistency was summarized using chance-corrected balanced accuracy and within-dataset ranks. Figure 5a presents the chance-corrected results across the six dataset entries together with their equal-dataset means. TopoSeqNet achieved the highest mean chance-corrected balanced accuracy (0.682), compared with 0.654 for the dataset-selected specific/classical reference, 0.650 for EEG Conformer, 0.648 for EEGNet, and 0.438 for the dataset-selected common reference. The dataset-selected specific/classical series draws on different paradigm-specific representations and specialized classical methods across datasets, providing a strong category-level reference against which to evaluate the consistency of the shared TopoSeqNet pipeline.

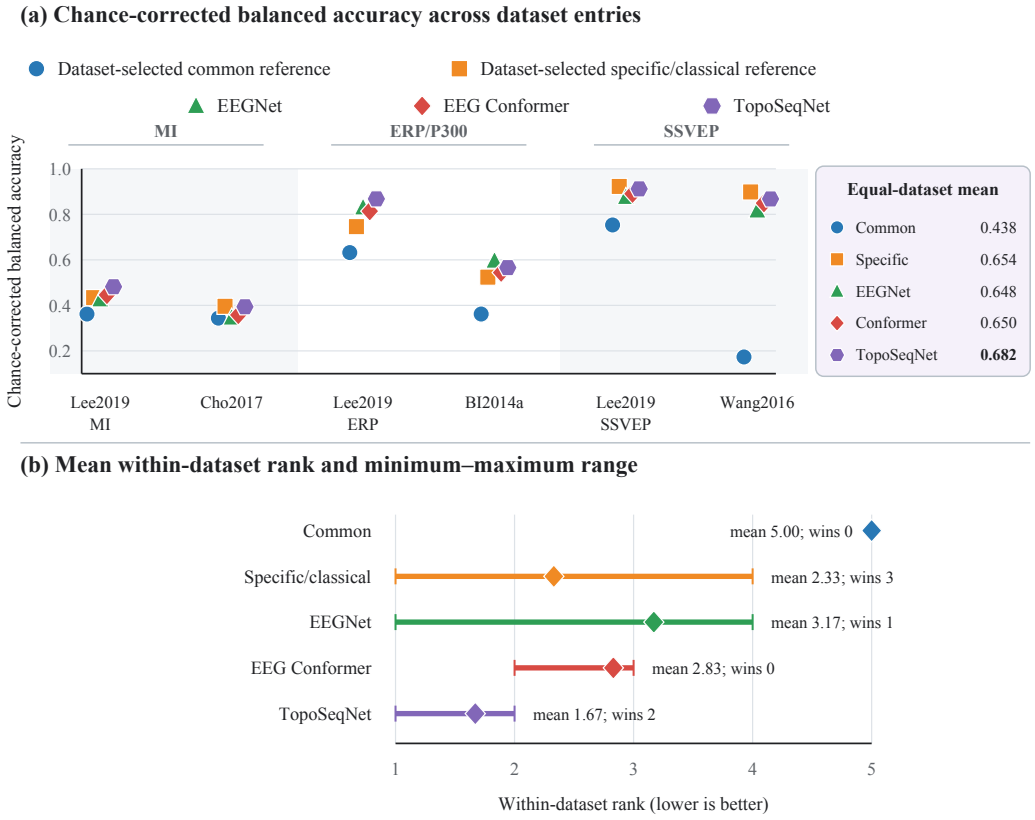


Figure 5. Cross-dataset consistency of the five principal comparison series. (**a**) Chance-corrected balanced accuracy across the six dataset entries; unconnected markers report descriptive within-dataset results, and the right panel reports equal-dataset means. (**b**) Mean and minimum-to-maximum range of within-dataset ranks; lower ranks indicate stronger performance, and annotations report mean rank and number of first-place dataset entries. Acquisition-source-weighted and leave-one-source-out summaries are reported in Table 8.

The complementary rank analysis in Figure 5b further demonstrates this consistency. TopoSeqNet achieved the best average rank (1.67) that remained within ranks 1–2 across all six dataset entries, and ranked first on two. The dataset-selected specific/classical reference had an average rank of 2.33 and a range of 1–4; EEG Conformer had an average rank of 2.83 and remained between ranks 2 and 3; and EEGNet had an average rank of 3.17 and a range of 1–4. The dataset-selected common reference ranked fifth on every dataset. Thus, TopoSeqNet combined the highest equal-dataset mean chance-corrected performance with the strongest overall rank profile.

Table 8. Robustness of the descriptive cross-dataset summaries to acquisition-source weighting and leave-one-acquisition-source-out analysis. Entries are mean chance-corrected balanced accuracy for the five principal comparison series; the equal-dataset summary with Wang2016 omitted is also reported to quantify its leverage. Bold values indicate the highest value within each summary row.

	Dataset-Selected References		Neural-Network Pipelines		
Descriptive Summary	**Common**	**Specific/Classical**	**EEGNet**	**EEG Conformer**	**Topo SeqNet**
Equal-dataset mean (six entries)	0.4378	0.6535	0.6484	0.6503	**0.6816**
Equal-dataset mean, Wang2016 omitted (five entries) [a]	0.4907	0.6045	0.6148	0.6105	**0.6444**
Acquisition-source-weighted mean (four sources) [b]	0.3654	0.6298	0.6173	0.6170	**0.6454**
OpenBMI omitted	0.2931	0.6062	0.5861	0.5837	**0.6092**
Cho2017 omitted	0.3726	0.7078	0.7077	0.7034	**0.7292**
BI2014a omitted	0.3666	0.6651	0.6244	0.6414	**0.6719**
Wang2016 omitted (source-weighted)	0.4295	0.5403	0.5509	0.5396	**0.5713**

[a] Excluding Wang2016 quantifies its leverage but does not isolate the separate effects of using 4 s rather than 5 s or of the Welch frequency resolution; those analyses were not rerun under controlled settings. [b] The three Lee2019 entries were first averaged within each principal series to form one OpenBMI acquisition-source value; OpenBMI, Cho2017, BI2014a, and Wang2016 then received equal weight. The final four rows repeat this source-weighted calculation after omitting the named source. All summaries are descriptive.

The sensitivity summaries in Table 8 support the same descriptive conclusion under alternative aggregation schemes. TopoSeqNet retained the highest mean chance-corrected balanced accuracy after equal weighting of the four acquisition sources (0.6454), after omitting Wang2016 from the equal-dataset summary (0.6444), and in each of the four leave-one-source-out summaries (0.5713–0.7292). These results show that its leading descriptive profile is preserved across dataset-level and acquisition-source-level summaries.

3.4. Participant-Level Statistical Comparisons

Figure 6 separates the two predefined neural-reference comparisons from the two exploratory comparisons involving dataset-selected targets. Holm correction was applied within each dataset across the same four-comparison family defined in Section 2.7; visual separation indicates the different status of the targets and does not change the multiplicity correction.

In the predefined comparisons, TopoSeqNet had significantly higher participant-level balanced accuracy than EEGNet on Lee2019_MI, Lee2019_ERP, Lee2019_SSVEP, and Wang2016. EEGNet had significantly higher accuracy on BI2014a, and the Cho2017 difference was not significant. Relative to EEG Conformer, TopoSeqNet had significantly higher accuracy on Lee2019_ERP and Lee2019_SSVEP; the remaining four differences were not significant. A non-significant result is not interpreted as evidence of equivalence.

Comparisons with the selected category references are exploratory because each target was chosen from the observed dataset-level means. Relative to the selected common reference, the mean TopoSeqNet difference was positive on all six datasets and Holm-significant on five. Relative to the selected specific/classical reference, the difference was significantly positive on Lee2019_ERP ($\Delta = 0.061$, 95% BCa CI $[0.049, 0.073]$, $p_{\mathrm{Holm}} < 0.001$) and BI2014a ($\Delta = 0.021$, 95% BCa CI $[0.004, 0.040]$, $p_{\mathrm{Holm}} = 0.0498$); the other four comparisons were not significant. These p-values and intervals are conditional on post hoc comparator selection and therefore support exploratory interpretation.

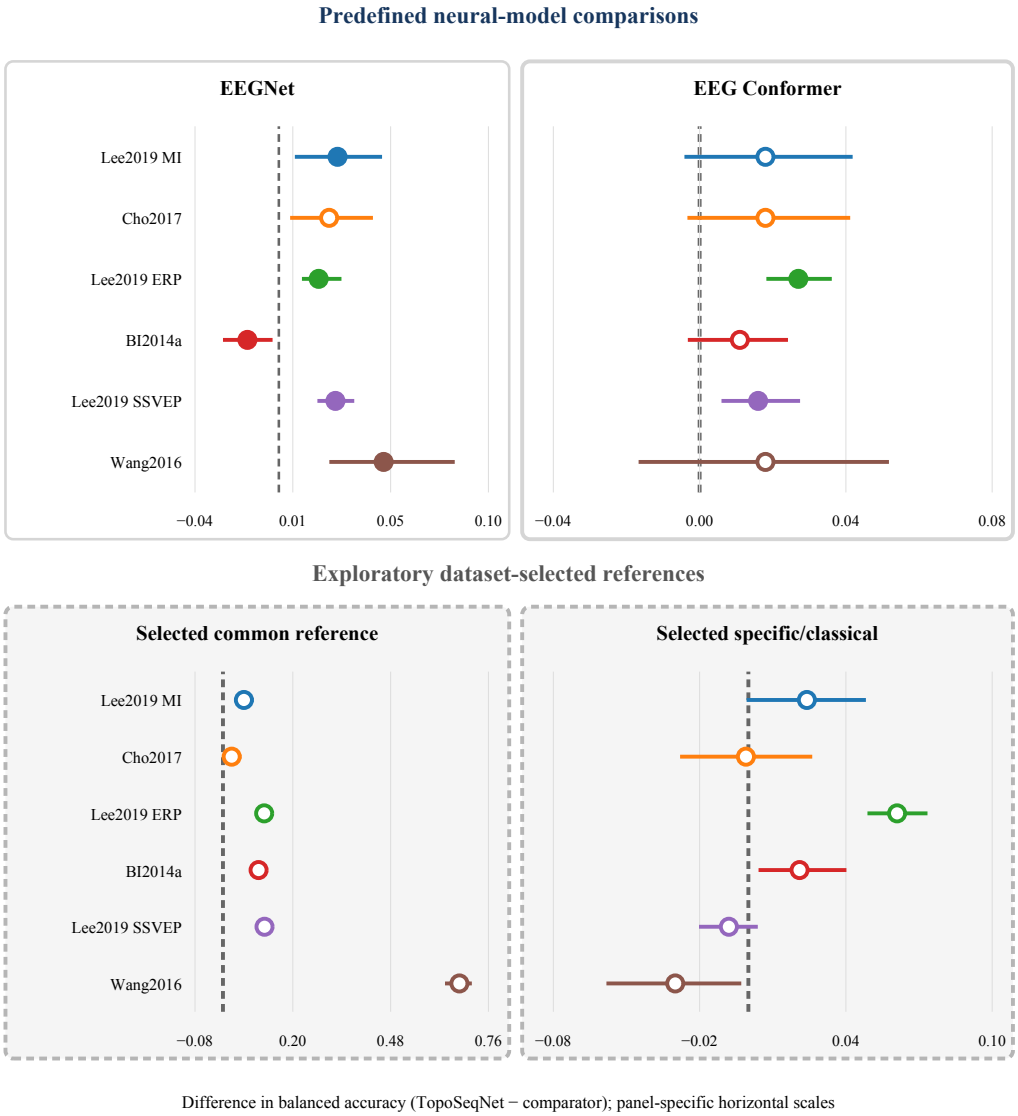


Figure 6. Participant-level paired balanced-accuracy differences between TopoSeqNet and the four comparison targets. Positive values favor TopoSeqNet; points show mean paired differences and bars show unadjusted 95% BCa confidence intervals. Predefined comparisons: EEGNet and EEG Conformer. Exploratory dataset-selected references: selected common and selected specific/classical pipelines, each chosen after inspection of the observed dataset-level means. Filled markers in the predefined panels indicate Holm-adjusted significance; exploratory markers remain open regardless of their adjusted *p*-values. Adjusted *p*-values and rank-biserial effect sizes are reported in Table 9.

Table 9. Participant-level paired comparisons between TopoSeqNet and the four reference series. Differences are TopoSeqNet minus the comparator. BCa: bias-corrected and accelerated bootstrap; p_H: permutation-test *p*-value after Holm correction within each dataset across the four headline comparisons; r_{rb}: paired rank-biserial correlation. The 95% BCa confidence intervals are unadjusted.

Status	Dataset/Comparator	*N*	Mean ΔBA (95% BCa CI)	p_H	r_{rb}
Predefined	Lee2019_MI/EEGNet	54	0.028 [0.008, 0.049]	0.0314	0.363
Predefined	Cho2017/EEGNet	52	0.024 [0.005, 0.045]	0.0814	0.316
Predefined	Lee2019_ERP/EEGNet	54	0.019 [0.011, 0.030]	0.00010	0.599
Predefined	BI2014a/EEGNet	64	−0.015 [−0.027, −0.003]	0.0476	−0.344
Predefined	Lee2019_SSVEP/EEGNet	54	0.027 [0.018, 0.036]	4.0×10^{-5}	0.779
Predefined	Wang2016/EEGNet	34	0.050 [0.024, 0.084]	0.00345	0.593

Table 9. *Cont.*

Status	Dataset/Comparator	N	Mean ΔBA (95% BCa CI)	p_H	r_{rb}
Predefined	Lee2019_MI/EEG Conformer	54	0.018 [−0.004, 0.042]	0.135	0.228
Predefined	Cho2017/EEG Conformer	52	0.018 [−0.003, 0.041]	0.252	0.205
Predefined	Lee2019_ERP/EEG Conformer	54	0.027 [0.018, 0.036]	4.0×10^{-5}	0.724
Predefined	BI2014a/EEG Conformer	64	0.011 [−0.003, 0.024]	0.127	0.279
Predefined	Lee2019_SSVEP/EEG Conformer	54	0.016 [0.006, 0.027]	0.00976	0.421
Predefined	Wang2016/EEG Conformer	34	0.018 [−0.017, 0.052]	0.316	0.200
Exploratory	Lee2019_MI/selected common	54	0.060 [0.032, 0.088]	0.000640	0.545
Exploratory	Cho2017/selected common	52	0.025 [−0.0005, 0.0530]	0.229	0.254
Exploratory	Lee2019_ERP/selected common	54	0.118 [0.103, 0.134]	4.0×10^{-5}	1.000
Exploratory	BI2014a/selected common	64	0.102 [0.080, 0.125]	4.0×10^{-5}	0.887
Exploratory	Lee2019_SSVEP/selected common	54	0.119 [0.097, 0.142]	4.0×10^{-5}	0.973
Exploratory	Wang2016/selected common	34	0.677 [0.636, 0.712]	4.0×10^{-5}	1.000
Exploratory	Lee2019_MI/selected specific/classical	54	0.024 [−0.001, 0.048]	0.122	0.289
Exploratory	Cho2017/selected specific/classical	52	−0.001 [−0.028, 0.026]	0.944	−0.007
Exploratory	Lee2019_ERP/selected specific/classical	54	0.061 [0.049, 0.073]	4.0×10^{-5}	0.934
Exploratory	BI2014a/selected specific/classical	64	0.021 [0.004, 0.040]	0.0498	0.263
Exploratory	Lee2019_SSVEP/selected specific/classical	54	−0.008 [−0.020, 0.004]	0.212	−0.195
Exploratory	Wang2016/selected specific/classical	34	−0.030 [−0.058, −0.003]	0.0867	−0.395

Note. The common and specific/classical comparators were selected according to their observed dataset-wise mean performance. Their comparisons are exploratory and conditional on post-selection; confirmatory interpretation is reserved for the predefined EEGNet and EEG Conformer comparisons.

4. Discussion

Using fixed configurations within one MOABB-based LOSO framework, this study demonstrates that EEG decoding performance depends strongly on the representation–classifier–dataset combination. Against this variable background, TopoSeqNet displayed a consistent cross-paradigm profile: the same topomap-sequence rule and architecture template remained competitive across all six dataset entries. The contribution is the empirical evaluation of cross-paradigm consistency as a design criterion alongside dataset-specific peak performance.

4.1. Dataset Dependence of Conventional EEG Pipelines

Figure 2 shows that the strongest feature–classifier combination changed across datasets, suggesting that many dataset-level winners reflect alignment among representation, classifier bias, and dataset structure rather than general superiority. Common_all supplied the dataset-selected common reference on three dataset entries, log-bandpower on two, and time statistics on one; the selected classifier also varied across entries, and the specified Wang2016 common-feature configuration yielded 0.194 balanced accuracy. Figure 3 further shows how effectively FBCSP, xDAWN, and FBCCA exploit task-matched assumptions. Together, Figures 2 and 3 quantify the feature–classifier selection challenge and provide the reference point for evaluating a shared cross-paradigm design.

4.2. Cross-Paradigm Consistency of the Topomap-Sequence Pipeline

The central result, summarized in Table 7 and Figure 5, is that the TopoSeqNet pipeline, using the same representation-construction rule and architecture template, ranked first or second across all six dataset entries. The task-informed references provide complementary

performance context through motor-band definitions, ERP windows, stimulus-frequency templates, and supervised spatial filters.

The representation is intentionally structured to retain electrode geometry in each frame and temporal order across frames. This provides a physiology-informed basis for accommodating lateralized rhythms, transient responses, and sustained frequency-locked activity within one input format. The performance results motivate targeted attribution and frequency-sensitive analyses that can determine how the fitted models use these dimensions in each paradigm.

The comparison with EEG Conformer adds an informative reference: across the six dataset entries, the topomap-sequence pipeline achieved a stronger rank profile than the raw-EEG convolution–attention model, with mean ranks of 1.67 and 2.83, respectively, and rank ranges of 1–2 and 2–3. This result motivates a controlled study of how input organization and architecture interact.

4.3. Paradigm- and Dataset-Specific Interpretation

The strongest advantage of TopoSeqNet over the dataset-selected specific/classical reference appeared for ERP/P300, particularly for the high-density OpenBMI dataset, as shown by the participant-level differences in Figure 6. The P300 component combines a temporal deflection with an evolving centro-parietal scalp distribution. Topomap sequences organize these temporal and spatial dimensions jointly, offering a plausible basis for the observed margin over the evaluated temporal-window and xDAWN references. Attribution analysis can test this interpretation directly.

The BI2014a result provides an informative limitation. Its 16-channel dry-electrode montage supplies much less spatial information for interpolation than the denser montages in the higher-density datasets. Constructing a dense grid from sparse electrodes necessarily introduces a greater proportion of estimated values, whereas EEGNet operates directly on the recorded channel–time matrix. The BI2014a result therefore highlights montage density and coverage as design factors for further study and motivates montage-aware interpolation and graph-based spatial modeling for low-density arrays.

For SSVEP, direct FBCCA remained numerically strongest because it incorporates the known stimulus frequencies, harmonics, and filter banks. TopoSeqNet nevertheless remained close on both the 4-class and 40-class datasets, while using the same learned representation strategy applied to MI and ERP/P300. The participant-level comparisons in Figure 6 quantify these differences; direct FBCCA's numerical advantage illustrates the value of exact frequency priors when the stimulation design is known. For Wang2016, the 0.2 Hz class spacing, approximately 1 Hz Welch resolution, and 4 s versus 5 s input durations define the context for interpreting the common-feature result.

The MI results followed a similar pattern. FBCSP remained highly competitive because its filters are optimized specifically for sensorimotor variance. TopoSeqNet achieved a mean balanced accuracy 0.024 higher than FBCSP on Lee2019_MI and 0.001 lower on Cho2017 while using the same spatiotemporal input construction employed for ERP/P300 and SSVEP. This pattern provides complementary evidence that one learned pipeline can remain competitive alongside a highly specialized MI method.

4.4. Interpretation of the Statistical Evidence

Figure 6 provides participant-level context for the cross-dataset summary. TopoSeqNet exceeded the dataset-selected common reference on all six datasets, with Holm-adjusted significance on five, while its largest differences from the dataset-selected specific/classical reference occurred for ERP/P300. Across the predefined neural comparisons, participant-level differences favored TopoSeqNet on several datasets. Because TopoSeqNet combines

topomap inputs with a CNN encoder containing squeeze-and-excitation blocks, followed by a Transformer, these findings characterize the integrated pipeline and motivate controlled component analysis of the representation–architecture interaction.

Comparisons with the dataset-selected common and specific/classical references remain exploratory because those references were selected from their observed LOSO means. The descriptive summaries in Figure 5 complement these selected-reference tests: TopoSeqNet combined the highest equal-dataset mean chance-corrected balanced accuracy with the lowest average rank and remained within ranks 1–2 while applying one fixed representation-construction rule and architecture template.

4.5. Scope, Limitations, and Future Work

Handcrafted and specialized methods remain valuable when the paradigm and acquisition conditions are known; FBCSP, xDAWN, and FBCCA exemplify the strength of well-matched signal priors. Across the heterogeneous benchmark settings represented here, the results highlight the practical value of a shared pipeline when the leading feature–classifier combination is unknown in advance.

The current evidence spans six dataset entries, three paradigms, and four acquisition sources, providing a clear foundation for broader task and acquisition coverage. Topomap quality depends on electrode locations and may degrade for sparse or irregular montages. The present evaluation targets subject generalization within each dataset. Fixed hyperparameters were intentionally used to maintain a common benchmark configuration, so dataset-specific tuning could alter individual rankings. From a deployment perspective, the topomap-sequence pipeline includes map-construction and sequence-processing stages, whereas raw-EEG models operate directly on channel-time arrays. A standardized deployment study can complement the predictive benchmark with calibration, parameter count, map-construction overhead, training time, inference latency, and memory profiles measured on common hardware.

Future work should examine additional datasets, low-density wearable systems, and other EEG paradigms. Direct cross-dataset transfer would provide a stricter test of representation invariance. Montage-aware interpolation, graph-based spatial modeling, and self-supervised pretraining may improve robustness across electrode configurations, while attention or saliency analyses could clarify which scalp regions and temporal intervals drive the predictions. A focused sensitivity study across grid resolutions, temporal aggregation windows, and interpolation rules would quantify design robustness, especially for low-density montages.

5. Conclusions

Under a unified subject-independent evaluation across six MI, ERP/P300, and SSVEP dataset entries, the leading common handcrafted feature–classifier pairing varied with the dataset, while FBCSP, xDAWN, and FBCCA achieved strong task-specific results through well-matched signal priors. Against these references, the same prespecified TopoSeqNet pipeline ranked first or second on every dataset entry and retained the highest mean chance-corrected balanced accuracy under both equal-dataset and acquisition-source-weighted summaries, together with the lowest average rank among the five principal comparison series. Together, these findings support the practical value of an integrated shared topomap-sequence pipeline as a reusable representation–architecture design that remains competitive alongside specialized classical and raw-EEG neural methods. The component, transfer, interpretability, sensitivity, and computational analyses motivated by this benchmark provide a clear path for extending the evidence.

https://doi.org/10.3390/a19090806

Author Contributions: Conceptualization, H.A. and M.S.; methodology, H.A.; software, H.A. and M.S.; validation, H.A., M.S. and L.M.; formal analysis, H.A. and M.S.; investigation, H.A.; resources, H.A., M.S., and L.M.; data curation, H.A. and M.S.; writing—original draft preparation, H.A.; writing—review and editing, H.A., M.S. and L.M.; visualization, H.A.; supervision, L.M.; project administration, H.A. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Regione Piemonte and co-funded by the European Union through the PR FESR Piemonte 2021–2027 programme, SWIch 2024 call, project BCI-SMILE (Brain-Computer Interface for Social inclusion, Mobility, Independence, and Life Enhancement; application no. 207145, grant no. E13D24000840006).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this study are publicly available through the MOABB framework and their respective repositories. The study repository, containing preprocessing, feature-extraction, model-training, evaluation, and figure-generation materials, is openly available at https://github.com/HosseinAhmadi63/Paradigm-Agnostic-EEG-Decoding (accessed on 4 September 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Abiri, R.; Borhani, S.; Sellers, E.W.; Jiang, Y.; Zhao, X. A Comprehensive Review of EEG-Based Brain–Computer Interface Paradigms. *J. Neural Eng.* **2019**, *16*, 011001. [CrossRef] [PubMed]
2. Ahmadi, H.; Impagnatiello, M.; Mesin, L. Semantic Latent Geometry Reveals Imagination–Perception Structure in EEG. *Appl. Sci.* **2026**, *16*, 661. [CrossRef]
3. Ahmadi, H.; Mesin, L. Universal Semantic Feature Extraction from EEG Signals: A Task-Independent Framework. *J. Neural Eng.* **2025**, *22*, 036003. [CrossRef] [PubMed]
4. Ang, K.K.; Chin, Z.Y.; Zhang, H.; Guan, C. Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008 ; pp. 2391–2398. [CrossRef]
5. Rivet, B.; Souloumiac, A.; Attina, V.; Gibert, G. xDAWN Algorithm to Enhance Evoked Potentials: Application to Brain-Computer Interface. *IEEE Trans. Biomed. Eng.* **2009**, *56*, 2035–2043. [CrossRef] [PubMed]
6. Lin, Z.; Zhang, C.; Wu, W.; Gao, X. Frequency Recognition Based on Canonical Correlation Analysis for SSVEP-Based BCIs. *IEEE Trans. Biomed. Eng.* **2007**, *54*, 1172–1176. [CrossRef] [PubMed]
7. Chen, X.; Wang, Y.; Gao, S.; Jung, T.P.; Gao, X. Filter Bank Canonical Correlation Analysis for Implementing a High-Speed SSVEP-Based Brain-Computer Interface. *J. Neural Eng.* **2015**, *12*, 046008. [CrossRef] [PubMed]
8. Jayaram, V.; Barachant, A. MOABB: Trustworthy algorithm benchmarking for BCIs. *J. Neural Eng.* **2018**, *15*, 066011. [CrossRef] [PubMed]
9. Aristimunha, B.; Carrara, I.; Guetschel, P.; Sedlar, S.; Rodrigues, P.; Sosulski, J.; Narayanan, D.; Bjareholt, E.; Barthelemy, Q.; Schirrmeister, R.T.; et al. *Mother of all BCI Benchmarks*, Version v1.7.2; Zenodo: Geneva, Switzerland, 2026. [CrossRef]
10. Ahmadi, H.; Mesin, L. Enhancing Motor Imagery Electroencephalography Classification with a Correlation-Optimized Weighted Stacking Ensemble Model. *Electronics* **2024**, *13*, 1033. [CrossRef]
11. Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. *Hum. Brain Mapp.* **2017**, *38*, 5391–5420. [CrossRef] [PubMed]
12. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A Compact Convolutional Neural Network for EEG-Based Brain–Computer Interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [CrossRef] [PubMed]
13. Song, Y.; Zheng, Q.; Liu, B.; Gao, X. EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2023**, *31*, 710–719. [CrossRef] [PubMed]
14. Rodrigues, P.L.C.; Jutten, C.; Congedo, M. Riemannian Procrustes Analysis: Transfer Learning for Brain–Computer Interfaces. *IEEE Trans. Biomed. Eng.* **2019**, *66*, 2390–2401. [CrossRef] [PubMed]
15. He, H.; Wu, D. Transfer Learning for Brain–Computer Interfaces: A Euclidean Space Data Alignment Approach. *IEEE Trans. Biomed. Eng.* **2020**, *67*, 399–410. [CrossRef] [PubMed]

16. Jiang, W.B.; Zhao, L.M.; Lu, B.L. Large Brain Model for Learning Generic Representations with Tremendous EEG Data in BCI. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
17. Ahmadi, H.; Santamaría-Vázquez, E.; Mesin, L.; Hornero, R. Semantic-Aware Decoding of Covert Inner Speech: A Multimodal EEG–EMG–Audio Framework. *Biomed. Signal Process. Control* **2026**, accepted for publication. [CrossRef]
18. Lee, M.H.; Kwon, O.Y.; Kim, Y.J.; Kim, H.K.; Lee, Y.E.; Williamson, J.; Fazli, S.; Lee, S.W. EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy. *GigaScience* **2019**, *8*, giz002. [CrossRef] [PubMed]
19. Cho, H.; Ahn, M.; Ahn, S.; Kwon, M.; Jun, S.C. EEG datasets for motor imagery brain–computer interface. *GigaScience* **2017**, *6*, gix034. [CrossRef] [PubMed]
20. Korczowski, L.; Ostaschenko, E.; Andreev, A.; Cattan, G.; Coelho Rodrigues, P.L.; Gautheret, V.; Congedo, M. *Brain Invaders Calibration-Less P300-Based BCI Using Dry EEG Electrodes Dataset (bi2014a)*, Version v1; Research Report; GIPSA-lab; Zenodo: Geneva, Switzerland, 2019. [CrossRef]
21. Wang, Y.; Chen, X.; Gao, X.; Gao, S. A Benchmark Dataset for SSVEP-Based Brain–Computer Interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2017**, *25*, 1746–1752. [CrossRef] [PubMed]
22. Welch, P.D. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short, Modified Periodograms. *IEEE Trans. Audio Electroacoust.* **1967**, *15*, 70–73. [CrossRef]
23. Hjorth, B. EEG Analysis Based on Time-Domain Properties. *Electroencephalogr. Clin. Neurophysiol.* **1970**, *29*, 306–310. [CrossRef] [PubMed]
24. Inouye, T.; Shinosaki, K.; Sakamoto, H.; Toi, S.; Ukai, S.; Iyama, A.; Katsuda, Y.; Hirano, M. Quantification of EEG Irregularity by Use of the Entropy of the Power Spectrum. *Electroencephalogr. Clin. Neurophysiol.* **1991**, *79*, 204–210. [CrossRef] [PubMed]
25. Duan, R.N.; Zhu, J.Y.; Lu, B.L. Differential Entropy Feature for EEG-Based Emotion Classification. In Proceedings of the 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), San Diego, CA, USA, 6–8 November 2013; pp. 81–84. [CrossRef]
26. Ledoit, O.; Wolf, M. A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. *J. Multivar. Anal.* **2004**, *88*, 365–411. [CrossRef]
27. Blankertz, B.; Lemm, S.; Treder, M.; Haufe, S.; Müller, K.R. Single-Trial Analysis and Classification of ERP Components: A Tutorial. *NeuroImage* **2011**, *56*, 814–825. [CrossRef] [PubMed]
28. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]
29. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
30. Ahmadi, H.; Mesin, L. Decoding Visual Imagination and Perception from EEG via Topomap Sequences. In Proceedings of the 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark, 14–17 July 2025; pp. 1–7. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.