SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V13I10P106

Volume 13 Issue 10, 62-82, October 2026
© 2026 Seventh Sense Research Group®

Original Article

Nonlinear Complexity Features Add Nothing beyond the Power Spectrum in Motor-Imagery BCI: A Surrogate-Controlled Reassessment

Manoj Kumar Mukul[1], Mukesh Kumar Ojha[2], Raj Kishore Prasad[3]

[1]*Department of Electronic Science, University of Delhi, South Campus, New Delhi, India.*
[2]*Greater Noida Institute of Technology (GNIOT), Greater Noida, Uttar Pradesh, India.*
[3]*Patna University, Patna, Bihar, India.*

[1]*Corresponding Author: manojkumarmukul@south.du.ac.in*

Received: 26 August 2026 Revised: 11 September 2026 Accepted: 24 September 2026 Published: 28 October 2026

Abstract *- A significant amount of research in the field of Motor-Imagery (MI) brain–computer interfaces have shown that nonlinear and higher-order features-such as fractal dimensions, entropies, detrended fluctuation analysis, variational-mode complexity, Phase–Amplitude Coupling (PAC), and bispectrum/bicoherence-enhance the decoding of sensorimotor rhythms. However, the past studies have drawn their conclusions based on within-session cross-validation, per-subject tuning, fusion over a spatial–spectral baseline, and the absence of a surrogate control. The present study poses a more focused and testable question: do these features contain information that goes beyond that provided by the power spectrum when assessed fairly? Using the BCI Competition IV datasets 2a and 2b (18 subjects) under the official independent-session protocol, a single matched shrinkage-LDA classifier is applied to decode fifteen families of features, each family is subjected to an Iterative Amplitude-Adjusted Fourier Transform (IAAFT) surrogate test, and all the families of tests are corrected for multiple comparisons. The results are clear. Nonlinear-7 achieves 0.576 compared to 0.691 for plain band power on dataset 2a and 0.668 compared to 0.735 on dataset 2b; when compared to a filter-bank CSP and Riemannian ensemble, the additional value of the surrogate-corrected nonlinear and PAC candidates is only +0.0000 Cohen's κ, and mutual-information selection includes them in 0.0% of the feature slots even though they make up 33.3% of the candidate pool. The higher-order spectra make the mechanism evident: across all 18 subjects, the magnitude of the bispectrum is statistically no different from that of band power (0.715 versus 0.721, p=0.59) and the accuracy on a per-subject basis correlates with band-power accuracy at r=0.958; in contrast, the normalized bicoherence, which removes that power scaling, achieves a decoding accuracy of 0.546-significantly lower than band power (p=3.5×10−6)-and falls below the level expected from its spectrally matched null (z=−0.16). A montage factorial that separates electrode count from spatial derivation shows that the apparent competitiveness of the complexity features on low-channel montages is not due to spatial focusing. This indicates the robustness of the results: when the number of channels, the number of electrodes, and the spatial derivation are kept the same across datasets, the ordering between nonlinear and band power remains unchanged, and the accuracy of the nonlinear features correlates with that of band power across the entire montage ladder (r=0.77 on 2a, r=0.84 on 2b). The results do not mean that all useful features disappear after controlling for the spectral information. In 3 of the 4 dataset-and-descriptor combinations, a small amount of information remains above chance even after correcting for false discoveries. However, this leftover effect is very small-only 4 to 6 percentage points above chance. By comparison, band power is 19 to 24 points above chance. And when these remaining features are added to the best-performing pipeline, they provide no additional benefit. This paper reports that nonlinear complexity features do not carry any usable information beyond the power spectrum for MI decoding, and the authors make available a reusable surrogate redundancy test that any candidate feature should pass before a nonlinear gain is claimed.*

Keywords *- Motor imagery, Brain-Computer Interface, Nonlinear features, Power spectrum, Surrogate data, IAAFT.*

1. Introduction

Since the first BCI meeting in 1999, research on developing a Brain–Computer Interface (BCI) inference system based on movement imagery has been a major focus. The key element of such a BCI inference system is feature extraction and selection. It is established that a pattern is a unique vector derived from the large amount of incoming EEG signal from the brain in response to an external stimulus; this vector is referred to as a feature. The majority of the features employed in the development of the BCI inference system are

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

based on linear system theory, this theory assuming that the EEG time series data conform to a linear system. Sensorimotor-rhythm motor imagery is the main paradigm used in non-invasive BCI. Its neurophysiological signature is well known in that imagining the movement of a limb leads to contralateral Event-Related Desynchronization (ERD) of the mu (8–13 Hz) and beta (13–30 Hz) rhythms in the area of the hand, together with corresponding ipsilateral synchronization [1]. This is a strong phenomenon, and the two feature representations most commonly used in practical systems, namely Common Spatial Patterns (CSP) [2] and band power, are, in essence, estimators of the spatially filtered spectral power.

Yet there is a strong biophysical argument for going beyond spectral analysis. It is a fact that the spectral analysis is based on the assumption of a linear, stationary, stochastic generator, but the cortex is in no way like that: individual neurons have sharp firing thresholds and input-output curves that saturate, and the populations they make up are connected by recurrent excitation and inhibition; as a result, the resulting mesoscopic dynamics are highly nonlinear.

Since the mid-1980s, physicists and neuroscientists have wondered whether the mathematics of nonlinear dynamics and deterministic chaos-developed for turbulence, lasers, and celestial mechanics-might also be applicable to the brain [3-5]. This line of research has given rise to a set of descriptors that are now standard in EEG analysis: the correlation dimension, the Lyapunov exponents, the waveform fractal dimensions, detrended fluctuation analysis, and entropy measures.

Early ambition in the form of low-dimensional deterministic brain chaos could not withstand rigorous methodological examination. Correlation dimension estimates, which seemed to reach a saturation point, were found to be reproducible from linearly filtered noise, as a result of which the field arrived at a more stable conclusion: although the EEG is nonlinear, it is not convincingly low-dimensional, and these measures should best be interpreted as comparative indicators of brain state rather than as estimates of the attractor's geometry [6, 7].

The point has not yet been widely accepted in the relevant MI-BCI literature, in which a continuous series of studies have stated that complexity measures lead to improved decoding. The explanation usually given is one that is plausible from a physiological point of view, since as a rhythm desynchronizes, the signal becomes more irregular and the fractal dimension and entropy therefore increase in the active hemisphere.

Yet this argument has a problem that is obvious. If complexity rises because mu power decreases, then complexity is just a rephrasing of power, and a direct estimate of power should at least perform as well. It is therefore an empirical question of a particular kind whether the nonlinear feature contributes anything, namely, whether it carries information beyond that of the power spectrum?

It is found in the literature survey that three methodological patterns recur across the applied literature and prevent the question from being answered:

1) Within-session cross-validation: Trials from one session share electrode placement, impedance, and subject state; random folds place neighbouring trials in training and test.
2) Fusion reporting: The nonlinear feature is concatenated onto a spatial-spectral pipeline, and the system accuracy is reported; therefore, the pipeline's contribution is credited to the feature.
3) No surrogate control: Nothing establishes that the measured quantity reflects nonlinearity rather than the spectral changes that accompany the ERD.

Research gap: Even though nonlinear and higher-order features are widely reported to improve motor-imagery decoding, no previous study has separated out their contribution from the power spectrum using a leakage-free independent-session protocol together with a matched classifier, a spectrally matched surrogate control, and corrections for the number of descriptors actually tested. Consequently, it remains unknown whether the reported gains reflect genuinely nonlinear information or merely a restatement of spectral power.

Closing this gap is the problem addressed in the present work. To that end, this study addressed the three methodological patterns directly. Fifteen families of features were tested using data from 18 people and two publicly available datasets. The study followed the official protocol for testing on independent sessions, used the same classifier for all experiments, tested each feature separately as well as in combination, and used an IAAFT surrogate test to check whether each feature family provided information beyond what is already contained in the signal's spectrum. This study makes three main contributions.

A reusable test for redundancy: It provides a surrogate-data method for determining whether a feature contains useful information that cannot be explained by the signal's spectrum alone. An explanation of why previous studies found improvements: The authors explain the underlying reasons for the reported performance gains, with the bispectrum and bicoherence analyses providing the clearest examples. A stronger reassessment of previous findings: The results are re-evaluated using five independent types of evidence, giving a more reliable picture of how useful these features actually are.

Novelty with respect to existing research: In contrast to the fractal-dimension [12], sub-band energy–entropy [13], and

entropy-based [14] motor-imagery studies, which report accuracy gains under within-session cross-validation, best-subject reporting, or fusion over a spatial–spectral pipeline, the present work is, to the authors' knowledge, the first motor-imagery study to (i) evaluate every nonlinear and higher-order family in isolation against a spectrally matched IAAFT surrogate null generated before decimation, (ii) require the surrogate-corrected candidates to compete inside the mutual-information feature pool of a state-of-the-art filter-bank CSP and Riemannian ensemble rather than being concatenated onto a weak baseline, and (iii) separate electrode count from spatial derivation through electrode-matched montage controls. None of these controls is present in the studies surveyed in Table 1, and together they change the interpretation of the previously reported gains.

2. Related Work and its Evaluation Gaps

Nonlinear measures do have a real clinical record [14]. In the case of epilepsy, Lehnertz and Elger [8, 9] found a decrease in the effective correlation dimension in the minutes leading up to a seizure, thus opening up the area of seizure prediction; on the Bonn benchmark [10], fractal and entropy features regularly achieve a seizure-detection accuracy of more than 95%.

Permutation entropy is used at the core of commercial depth-of-anaesthesia monitors since it tracks anesthetic concentration well [11]. They involve large and reliable changes that are measured between clearly distinct conditions, and there is no dispute regarding the descriptive value of complexity measures.

Within motor imagery itself, the literature is extensive. A recent review of 63 MI-BCI entropy studies published between 2019 and 2023 [14] found that sample, Shannon, fuzzy, permutation, and approximate entropy dominate the field, that most studies employ a single entropy feature, and that adding entropy features is typically credited with a gain of 8–10 percentage points in classification accuracy.

The fractal-dimension pipelines that combine estimators such as the Katz, Higuchi, and Petrosian dimensions report accuracies near 79% on BCI Competition IV-2a, above CSP and Riemannian baselines evaluated in the same within-session setting [12], and wavelet sub-band energy and entropy features of the mu and beta rhythms classified with LDA and SVM reach above 90% on the best subjects [13].

Almost without exception, however, these figures are obtained under the three evaluation patterns identified in Section 1-within-session folds, fusion or best-subject reporting, and the absence of any spectrally matched control-so the isolated, protocol-fair contribution of the nonlinear ingredient cannot be recovered from the published numbers alone. Establishing that contribution requires the controls introduced in Section 3.

Motor imagery presents a different issue: the contrast here lies between two imagined hands in the same subject, during the same session, and in the same cognitive state, and the discriminative signal consists of a lateralized, band-limited power modulation of a few decibels. Table 1 includes the representative MI reports (see also Figure 1).

Table 1. Representative motor imagery studies using nonlinear or higher-order features

Study/Family	Dataset	Subject	Validation	Reported Accuracy
Katz + Higuchi + Petrosian FD [12]	BCI IV-2a (4-class)	9	Within-session CV	~79% (> CSP 74%, Riemann 76%)
Sub-band energy + entropy [13]	BCI IV-2a	9	Best-subject reporting	Up to ~92% (best subject)
Bispectrum channel selection	BCI IV-2b	4	10-fold within-session	83.8% (vs. 69.4%)
-same method	BCI III-IVa	5	10-fold within-session	86.3% (vs. 82.9%)
-same method	BCI III-IIIa	3	10-fold within-session	77.8% (vs. 68.2%)
Higher-order statistics	BCI IV-2a	9	Within-session CV	High-70s–80s% [1]
PAC/cross-frequency + Riemann	BCI IV-2a (4-class)	9	Cross-session	78.2% (vs. FBCSP 69.9%)
This work-CSP baseline	BCI IV-2a (2-class)	9	Independent session (T→E)	0.78
This work-best nonlinear family	BCI IV-2a (2-class)	9	Independent session (T→E)	≤ 0.63

Figure 1 behind a paywall at the time of writing. Fusion and all-channel pipelines report the combined system, not the isolated feature. The accuracies across datasets are not directly comparable. Across the surveyed studies, the sequence is largely shared-acquisition, preprocessing, nonlinear/complexity feature extraction, optional feature selection, classification, and evaluation-but the specific choices at each step differ (Table 1). Evaluation is predominantly within-session cross-validation, frequently on the best subject or a two-class subset.

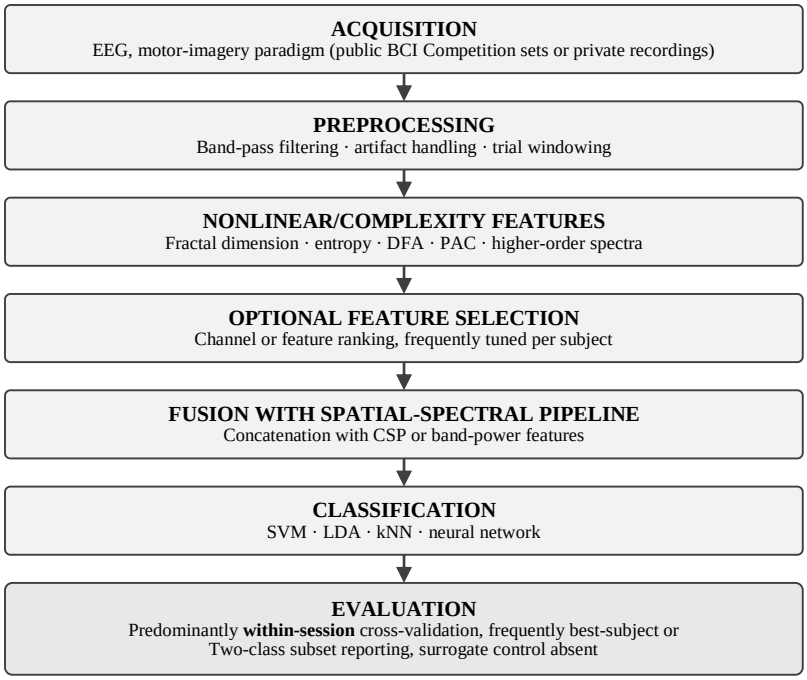


Fig. 1 Depicts the typical processing pipeline in prior work on nonlinear-feature motor-imagery classification

3. Materials and Methods

3.1. Data and Protocol

This work used the BCI Competition IV dataset 2a (22 monopolar channels, four classes, two sessions per subject, nine subjects, 250 Hz) and dataset 2b (three bipolar channels over C3, Cz, and C4, two classes, five sessions, nine subjects, 250 Hz). This work analyzed the two-class left-versus-right contrast common to both, which makes 2b the genuine low-channel counterpart to 2a and provides 18 subjects in total.

Three evaluation regimes were computed on the same features and classifier, so that the optimism attributable to the protocol could be measured directly rather than inferred across studies.

1) Validation-within-session 5-fold cross-validation in the training session.
2) Independent-session test-the official competition protocol: training on the T session, evaluation once on the E session using the released true labels.
3) Cross-subject leave-one-subject-out with Euclidean Alignment [30], which is label-free and therefore leaks no class information into the held-out subject.

The GDF files were read using BioSig. Trials were cut from the 0.5–2.5 s window after the cue onset. No trials were rejected beyond the competition's artefact handling. The three EOG channels were never used as features; to confirm that residual ocular activity was not doing the work, the entire analysis was re-run with EOG regressed out of the continuous EEG by least squares [15] before trial extraction (Section 4.11).

3.1.1. Montages: Channels are not Electrodes

Since the literature includes claims that low-channel montages are important, the montage is treated in this case as an explicit experimental factor rather than merely a preprocessing detail, and it is specified both in terms of electrodes and channels. As each bipolar channel requires two electrodes, a three-channel bipolar montage uses six sensors. Compared with a three-electrode monopolar montage, the additional sensors are automatically attributed to the derivation. Table 2 shows the montages used.

Unless otherwise stated, mono22 is used for decoding in 2a, and the full three-channel recording is used in 2b; for the C3/C4 comparisons in Sections 4.3 and 4.10, bip2 is employed. The signals are band-pass filtered between 8 and 30 Hz for the CSP, band-power and complexity families; the PAC and higher-order spectra use their own defined frequency bands.

3.2. Multiple Comparisons

All the families of tests described below-namely, the tests involving band power, the descriptors in the surrogate table, the descriptors in the lateralisation contrast, the montage contrasts, and the four residual tests-are subjected to correction using the Benjamini–Hochberg false-discovery-rate procedure within each family, and both the raw p value and the corrected q value are given. It is important to note that several effects which are merely significant do not survive.

3.3. Feature Families

All families were computed on the same channels and window and decoded by the same classifier.

CSP (spatial-spectral baseline). Three spatial-filter pairs from the standard CSP eigen-decomposition of the class covariance matrices and log-variance features.

Filter Bank CSP (FBCSP). CSP within sub-bands spanning 4–40 Hz with mutual information feature selection.

Band power (spectral baseline). Log-variance of the 8–30 Hz signal per channel-the simplest spectral feature.

Nonlinear-7 (Set A). Higuchi [20], Katz, and Petrosian fractal dimensions, the DFA exponent [23], and sample [21], permutation [22] and spectral entropy, per channel. This is the descriptor set used by the studies in Table 1 and is the primary set throughout.

Nonlinear-7 (Set B). A second, deliberately different set of seven: time-reversal asymmetry, permutation entropy of order 4, Higuchi FD with kmax=8, one-step prediction error, sample entropy, Lempel–Ziv complexity and correlation dimension. Set B is carried through the entire analysis in parallel with Set A. Its only purpose is to answer the objection that a negative result reflects an unlucky choice of descriptors: if the conclusion held for Set A alone, it could be dismissed that way, and if it holds for both, it cannot. Every table below that distinguishes the two reports, Set A first.

VMD-nonlinear. Variational mode decomposition [27] into K=4 narrowband modes per channel, with Higuchi FD, permutation entropy and log-power per mode. Modes are ordered by centre frequency before feature extraction; the ADMM does not preserve its initialisation order, and unordered modes would scramble the meaning of each feature column between trials.

PAC (direct). Tort modulation index [24] between mu-band phase (8–13 Hz, analytic signal) and gamma-band amplitude (30–45 Hz), computed by filtering and the Hilbert transform, within and across channels.

Bispectrum magnitude and bicoherence. [25] From Hann-windowed 128-point segments at 50% overlap, the authors apply the third-order product X(f1)X(f2)X∗(f1+f2). The raw magnitude |B(f1,f2)| defines the magnitude features, and the power-normalised bicoherence

$$b^2(f_1, f_2) = \frac{|\langle B \rangle|^2}{\langle |X(f_1)X(f_2)|^2 \rangle \langle |X(f_1+f_2)|^2 \rangle} \quad (1)$$

Which is zero for a linear Gaussian process, defines the bicoherence features. Both are band-averaged over mu–mu, mu–beta, and mu–gamma regions and summed over the 4–45 Hz plane.

Riemannian baselines. Minimum distance to mean and tangent-space projection with LDA [26]. Fusion. CSP+Nonlinear-7, CSP+PAC, and Nonlinear-7+PAC were concatenated to test whether any family adds to the spatial-spectral baseline.

3.4. Dynamical Characterisation

The method of phase-space reconstruction [19] involved the use of average mutual information for selecting the delay and employed a false nearest-neighbours embedding technique. The correlation dimension was calculated by applying the Grassberger–Procaccia [17] approach together with a Theiler window; the largest Lyapunov exponent was determined by means of Rosenstein's method [18]. Nonlinearity was evaluated using IAAFT surrogates and a time-reversal asymmetry statistic. The present branch provides the dynamical context against which the decoding results should be interpreted; it is not a decoding result itself (see Figure 2).

3.5. Matched Classifier

The classification of each family uses the same shrinkage-regularized LDA [29], the intensity of shrinkage being estimated analytically rather than being set beforehand, thereby ensuring that any differences between families are due to their characteristics and not to the classifier. It is not possible to use a single fixed value for both a six-dimensional CSP feature vector and a 253-dimensional tangent-space vector with the same number of trials. The decision to keep the classifier fixed is intentional, since it eliminates the most frequent confounding factor in the literature on nonlinear features, namely, the misconception that a better classifier is a result of the feature. The accuracy and Cohen's κ are reported.

3.6. The IAAFT Surrogate Redundancy Test

To determine whether a feature carries information beyond the power spectrum, the authors generate IAAFT surrogates of each trial. IAAFT iterates two projections to convergence: it imposes the exact Fourier magnitude spectrum of the data and then restores the exact amplitude histogram. The surrogate therefore matches the trial's power spectrum-hence its full linear autocorrelation-and its marginal distribution, while randomizing the phase. Everything that a linear process can explain is preserved; only the nonlinear temporal structure is destroyed. For a measure Q, one computes,

$$z = \frac{Q(\text{data}) - \langle Q(\text{surrogate}) \rangle}{\text{std}\left(Q(\text{surrogate})\right)} \quad (2)$$

With |z|>1.96 corresponding to two-sided significance at p≈0.05. Because the surrogate distribution of a strictly non-negative measure need not be Gaussian, the authors additionally report the rank-based fraction of trials that are individually significant, which makes no distributional assumption. For adaptive decompositions, the surrogate is re-decomposed; therefore, the null inherits no data-adaptive structure.

Two implementation details govern whether this test is honest, and both are easy to get wrong. Order of operations. The surrogate should be created at the original sampling rate and then subjected to the same decimation process as the data, rather than being generated from a series which has already been decimated. IAAFT is not capable of exactly matching both the amplitude spectrum and the value histogram at the same time, the degree of residual.

Table 2. Describes montages. Dataset 2b is bipolar as recorded (pairs ≈2.5 cm apart) and admits no monopolar derivation; dataset 2a is monopolar, and its bipolar pairs are derived across two rows of the 10–10 layout, roughly twice the 2b spacing. Each bipolar montage has an electrode-matched monopolar control built from exactly the same electrodes.

Name	Channels	Electrodes	Definition
mono22	22	22	all 2a channels, monopolar
mono6	6	6	FC3/FCz/FC4 + CP3/CPz/CP4, monopolar (control for bip3)
bip3	3	6	FC3−CP3, FCz−CPz, FC4−CP4 (2a); C3/Cz/C4 as recorded (2b)
mono4	4	4	FC3/FC4 + CP3/CP4, monopolar (control for bip2)
bip2	2	4	FC3−CP3, FC4−CP4 (2a); C3/C4 as recorded (2b)

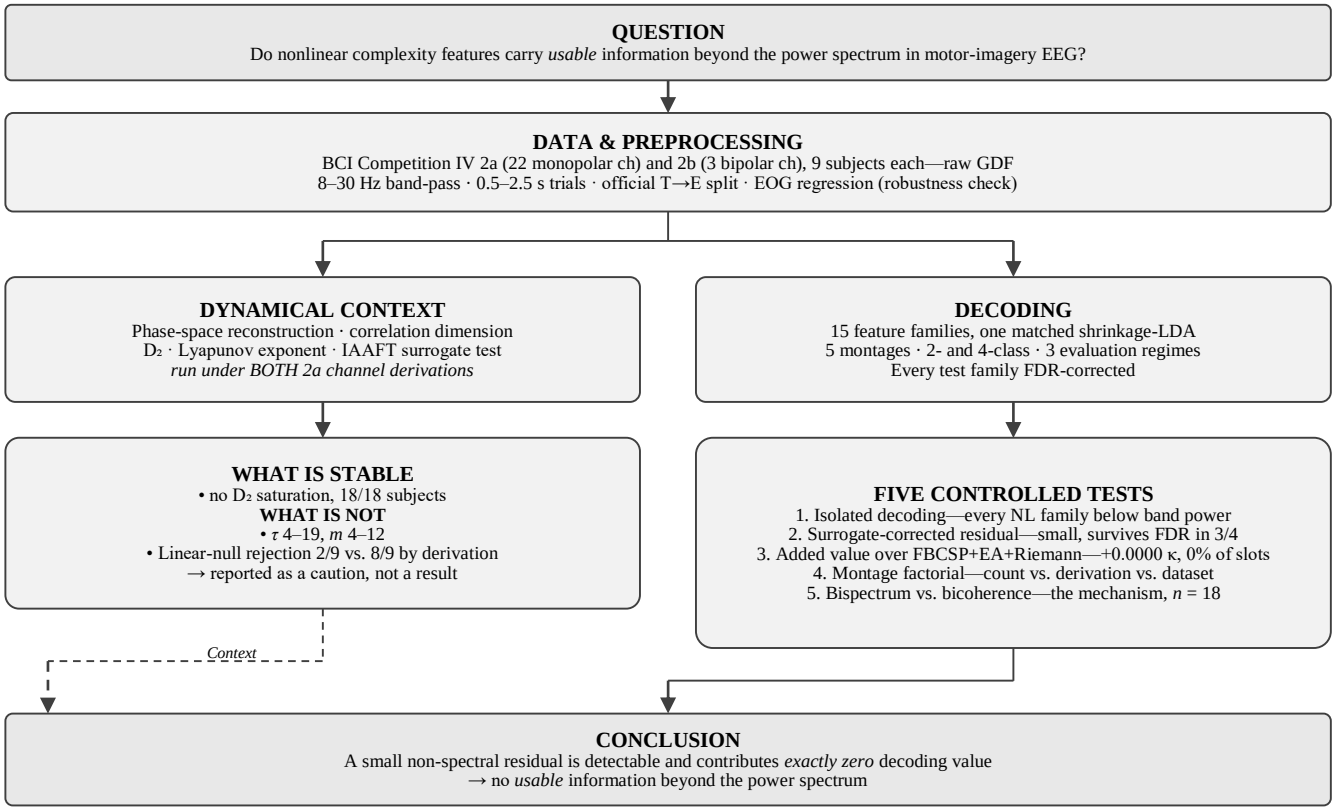


Fig. 2 Presents the flowchart of the analysis

The dynamical branch provides context and, in this revision, includes an explicit warning that only the lack of saturation in the correlation dimension is stable across different channel derivations. The decoding branch proceeds with the argument through five controlled tests, the most important of which is whether a candidate feature is ever chosen when competing within the feature pool of a state-of-the-art baseline. Mismatch depending on the Length of the Series: when the series has 125 samples, the relative spectral residual is 3.3%, 1.0% when it has 500 and 0.55% when it has 1000, with the number of iterations reaching a plateau of about 20. Therefore, generating the surrogate after decimation presents the algorithm with its worst possible scenario, and an imperfect null causes std (Qsurr) to increase, biasing the test in such a way as to lead to the conclusion that nothing has been found. Resolution of the rank test. When there are n surrogates the lowest possible two-sided rank p is 2 divided by (n+1); consequently, the statement "0% of the trials were significant at α=0.05" obtained from 19 surrogates is merely an arithmetic result and not a measured value; throughout the text $n \geq \lceil 2/\alpha \rceil - 1 = 39$ is used and the rank criterion is assessed using an explicit floating-point tolerance in order to make the test symmetric with respect to sign (see Figure 3).

The test has two modes. In feature-value mode, it determines whether the measured quantity exceeds a spectrally matched null; in decoding mode, the features are recalculated on surrogate trials, and the classifier is retrained to assess whether decoding accuracy remains after the nonlinear structure is destroyed. A feature which reflects true nonlinearity goes above the null in both cases, while a spectral proxy does not exceed it in either case.

4. Results

4.1. Contralateral Mu-ERD Confirms the Paradigm

The same datasets show the anticipated contralateral effects, which are much more pronounced when using the natively. Bipolar 2b montage. For the hemispheric lateralisation index (C3 − C4), the value is +17.4 in the case of left-hand imagery and −27.8 for right-hand imagery on the 2b montage, compared with +2.8 and −8.7 on the monopolar 2a montage.

EA = Euclidean Alignment.

Two-class left-vs-right subset, not the four-class competition ranking. All numbers are from a single program and a single window, so they are mutually consistent.The paradigm functions and physiological effects referred to in the complexity literature are present (see Figures 4 and 5).

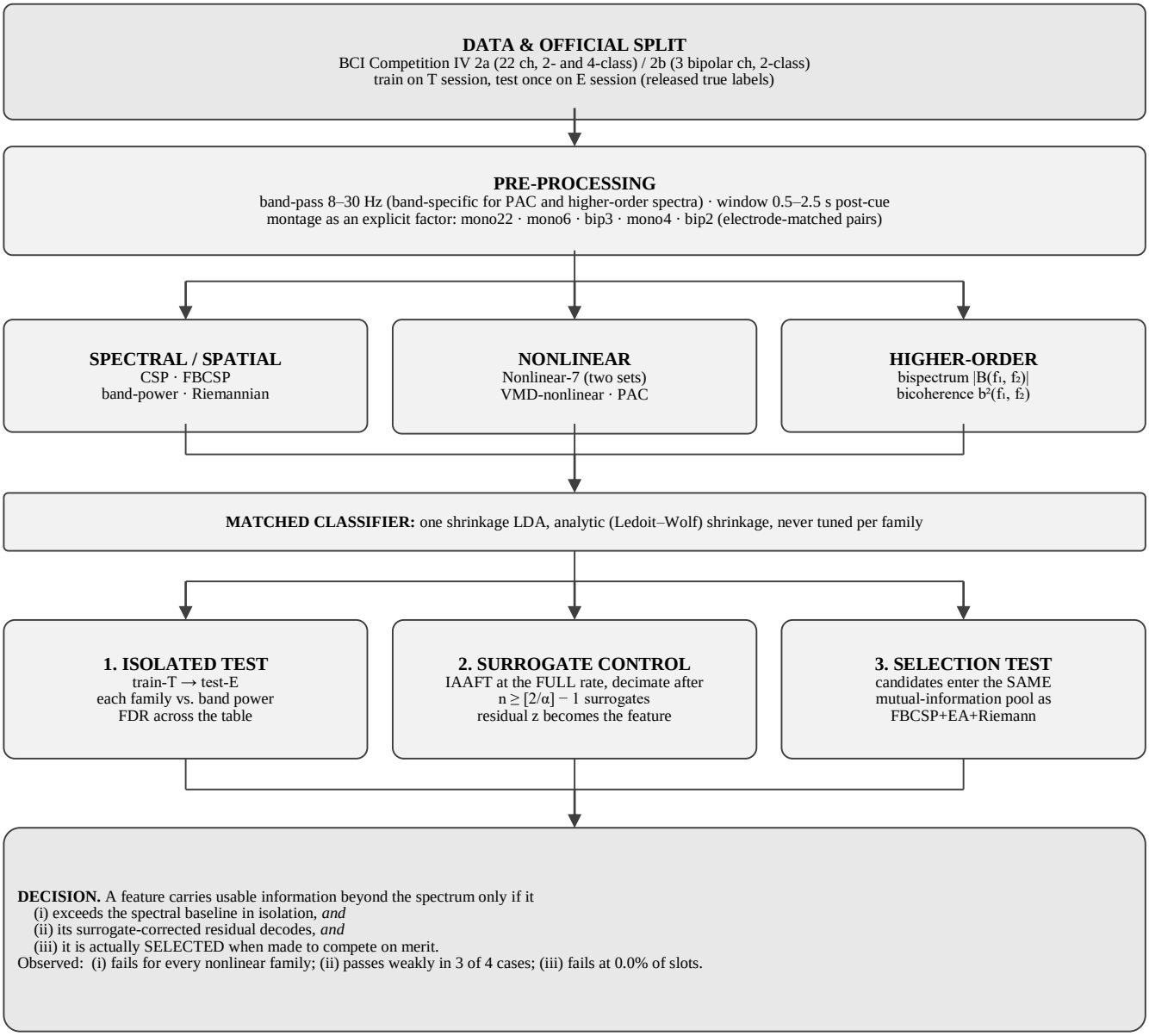


Fig. 3 The redundancy test as applied here. Two elements distinguish it from the usual surrogate check: the surrogate is generated at the original sampling rate and decimated afterwards (reversing the order biases the test towards finding nothing), and the candidate must survive a selection test inside the feature pool of a competent baseline rather than merely being concatenated onto a weak one. All three panels are drawn in TikZ and are editable in the manuscript source.

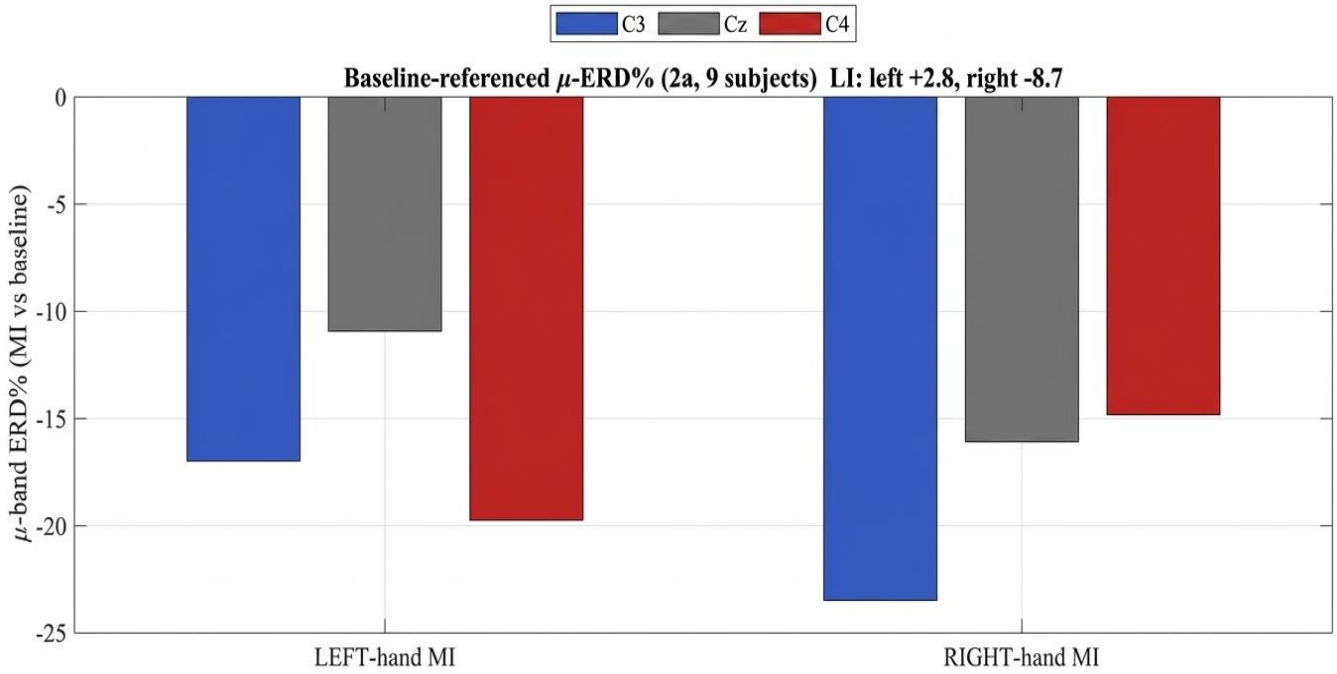


Fig. 4 Baseline-referenced mu-band (8–13 Hz) ERD% over C3, Cz, and C4 for each imagined hand, dataset 2a (nine subjects). The contralateral channel was least synchronized (left hand → C4 lowest; right hand → C3 lowest). Absolute values are positive because the pre-cue baseline follows the warning cue (see text).

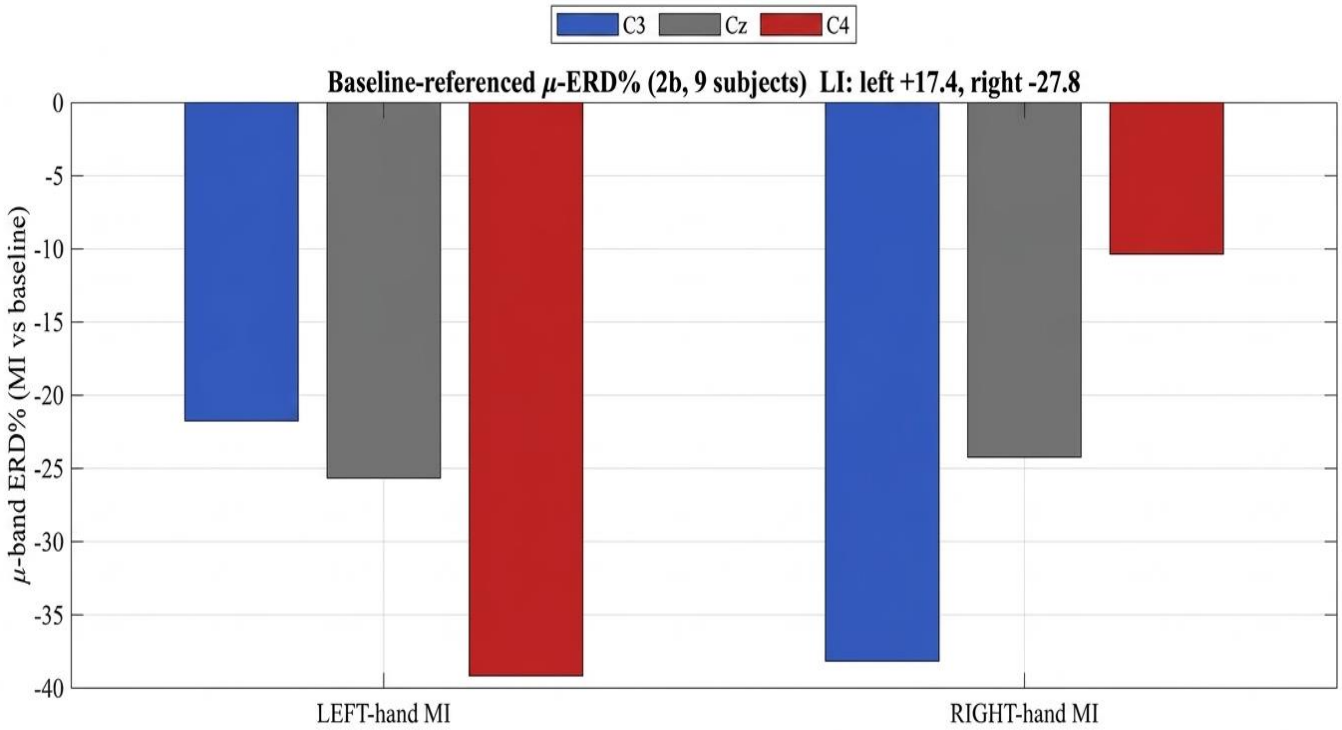


Fig. 5 Same as in Figure 4 for dataset 2b (bipolar montage), where contralateral lateralization is strong and bidirectional.

Table 3. Accuracy (Cohen's κ where computed) by regime, dataset, and feature set. Fusion = CSP + Nonlinear-7

Regime	2a CSP	2a Fusion	2b CSP	2b Fusion
Within-session CV	0.765	0.709	0.716	0.709
Official T→E (κ)	0.771 (0.542)	0.708 (0.417)	0.748 (0.495)	0.759 (0.517)
LOSO, EA (κ)	0.715 (0.430)	0.689 (0.378)	0.704 (0.407)	0.686 (0.372)

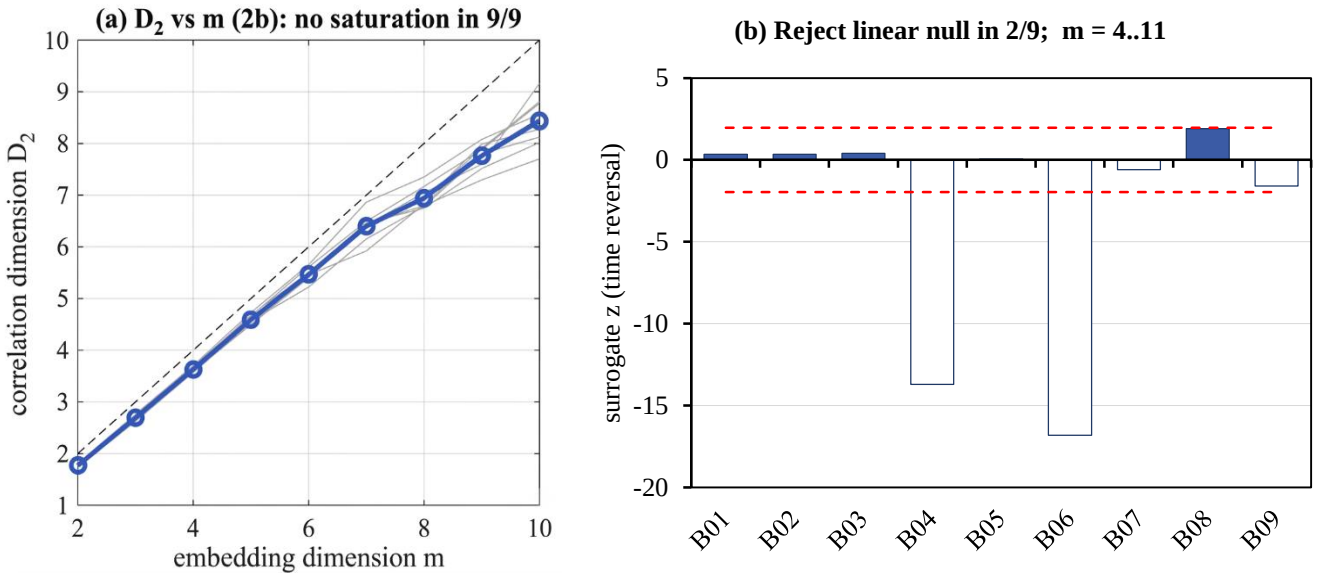


Fig. 6 Shows the pooled dynamical characterisation for dataset 2b. (a) For all nine subjects the correlation dimension D2 is plotted against the embedding dimension m and increases steadily without ever reaching a saturation point, and (b) The per-subject surrogate z value of the time-reversal statistic; the dashed lines indicate ±1.96. As in the case of Figure 7, the number of subjects lying outside those lines varies according to the channel derivation (Section 4.2).

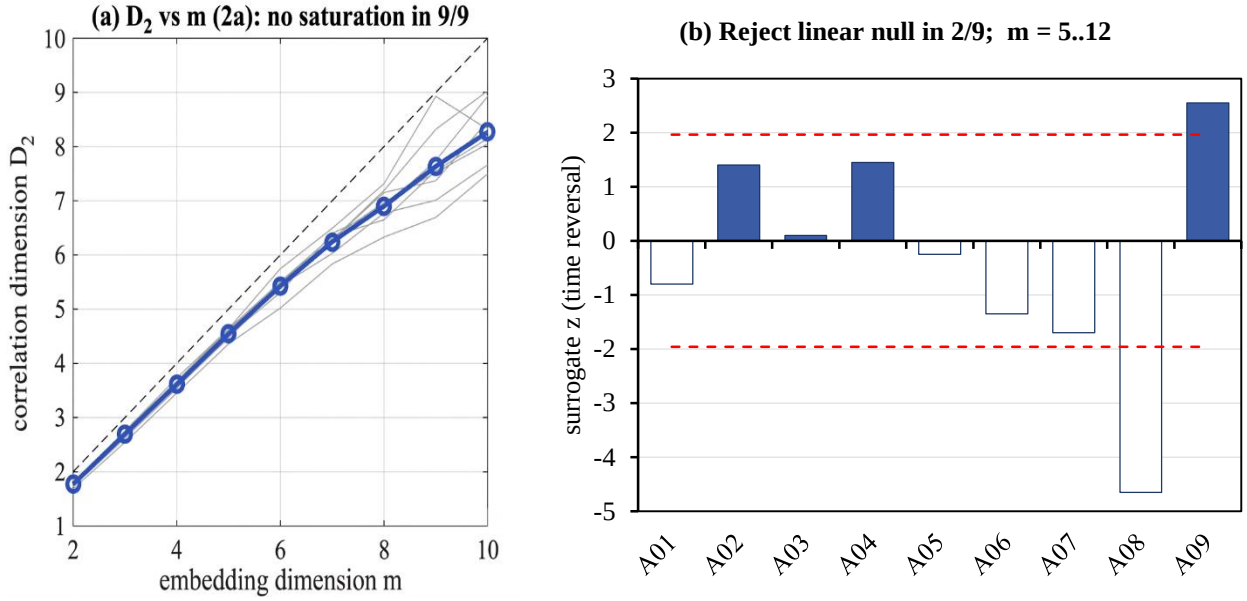


Fig. 7 Pooled dynamical characterization, dataset 2a. (a) Correlation dimension D_2 against embedding dimension m for all nine subjects: D_2 climbs steadily and never saturates, and (b) Per-Subject surrogate z of the time-reversal statistic; dashed lines mark ±1.96. The number of subjects beyond those lines depends on the channel derivation (Section 4.2), which is why it is not quoted as a result.

4.2. The Dynamical Picture does not Support Dimension-based Descriptors-and is not Stable Enough to Report as a Result

When the C3 segments are reconstructed, the correlation dimension increases steadily as the embedding dimension rises and never reaches a saturation point, this being the case in all 9 subjects for 2a and in all 9 subjects for 2b, reaching a value of D2≈8–9 by the time m=10. This behaviour is the characteristic signature of high-dimensional or stochastic dynamics rather than that of a low-dimensional attractor; it is robust since it depends on the shape of the relationship between D2 and m rather than on the actual value of D2, and it is the only dynamical conclusion stated here without reservation. The low-dimensional attractor, which would provide a basis for using dimension-based descriptors, is absent.

The remaining phase-space quantities are not sufficiently stable to report as findings, and this is stated rather than quoting the more convenient of the two answers. The selected

delay ranges from τ=4 to 19 samples across subjects and the false-nearest-neighbour embedding dimension from m=4 to 12; a single value of m=4 is obtained only under one particular set of FNN parameters (the R-criterion alone, threshold 0.05, returning the search ceiling when the criterion is never met).

It is also observed more seriously that the count of subjects rejecting the linear-stochastic null by the time-reversal statistic depends on a preprocessing choice that has no principled answer: on 2a, it is 2 of 9 under the derived bipolar FC−CP montage and 8 of 9 under the monopolar C3/C4 montage; on 2b, it is 2 of 9. The same data, the same estimator, and an arbitrary derivation swing the headline count from a small minority to a large majority.

This section is therefore treated as establishing one thing-no correlation-dimension saturation anywhere-and as a caution about the rest. Estimators designed for long, clean, low-dimensional records do not converge on 2-second epochs of scalp EEG, and a number that moves this much with an arbitrary montage choice should not be carrying an argument (see Figures 6 and 7).

4.3. Complexity Lateralisation does not Survive Correction

If the rhythm causes the signal to become more irregular, single-trial complexity should show lateralisation with respect to mu power, with greater values over the hemisphere opposite to the imagined hand. The authors applied all fourteen descriptors (Set A together with Set B) in a paired comparison between the contralateral and ipsilateral sides, using both channel derivations. No descriptors survive. With respect to dataset 2a, none of the 14 descriptors achieves a significance level of q<0.05 after applying the Benjamini–Hochberg correction under the derived bipolar montage, and similarly none do under the monopolar montage, even though the most notable nominal effects are observed in those cases. The largest uncorrected effects are permutation entropy of order 4 (t=−2.45, p=0.040, q=0.211) and correlation dimension (t=−2.04, p=0.075, q=0.211) in the case of the bipolar derivation, and permutation entropy (t=−2.34, p=0.048, q=0.378) with the monopolar derivation. When no correction is made, two or three of the descriptors have p<0.05, but when corrected for the 14 tests actually carried out, none do.

This is treated as a null result; although the physiological account which underlies the complexity features remains unrefuted-since the ERD in Section 4.1 is clear-the assertion that broadband single-trial complexity shows a detectable lateralisation is not substantiated when the multiplicity of the descriptor set is considered.

4.4. Decoding across Three Evaluation Regimes

Cohen's κ lands in the 0.38–0.54 range on the official protocol, 'fair' to 'moderate' on the Landis–Koch scale [28]. Cross-subject transfer is the hardest regime on both datasets, as it should be. The within-session and independent-session figures, however, are *not* ordered as the optimism argument predicts: on 2a, they are 0.765 and 0.771, and on 2b, 0.716 and 0.748, with the official test exceeding within-session cross-validation in both cases.

On 2b, the reason is structural-the evaluation sessions use online feedback, on which subjects perform better and more uniformly, whereas the training set folds in the tougher no-feedback screening runs. On 2a, the two regimes are simply within noise of one another at *n*=9 (Table 3).

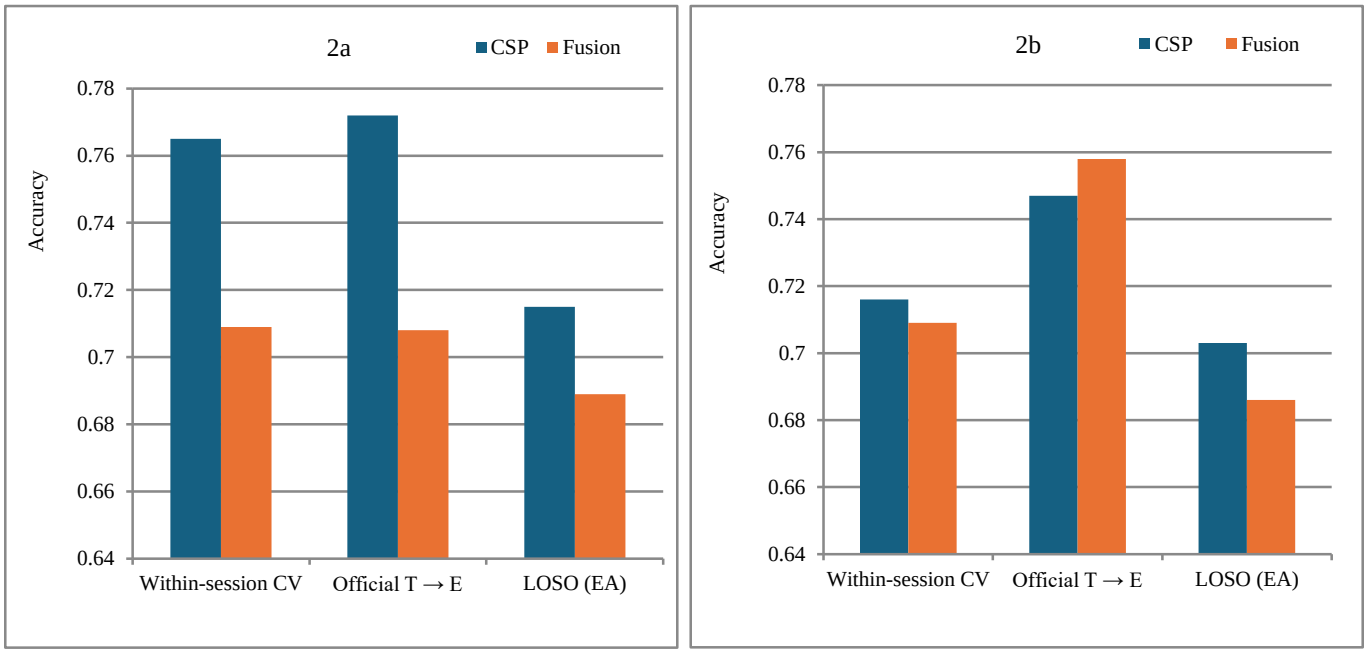


Fig. 8 Decoding across three regimes for both datasets (nine subjects each). Δ = fusion − CSP in percentage points

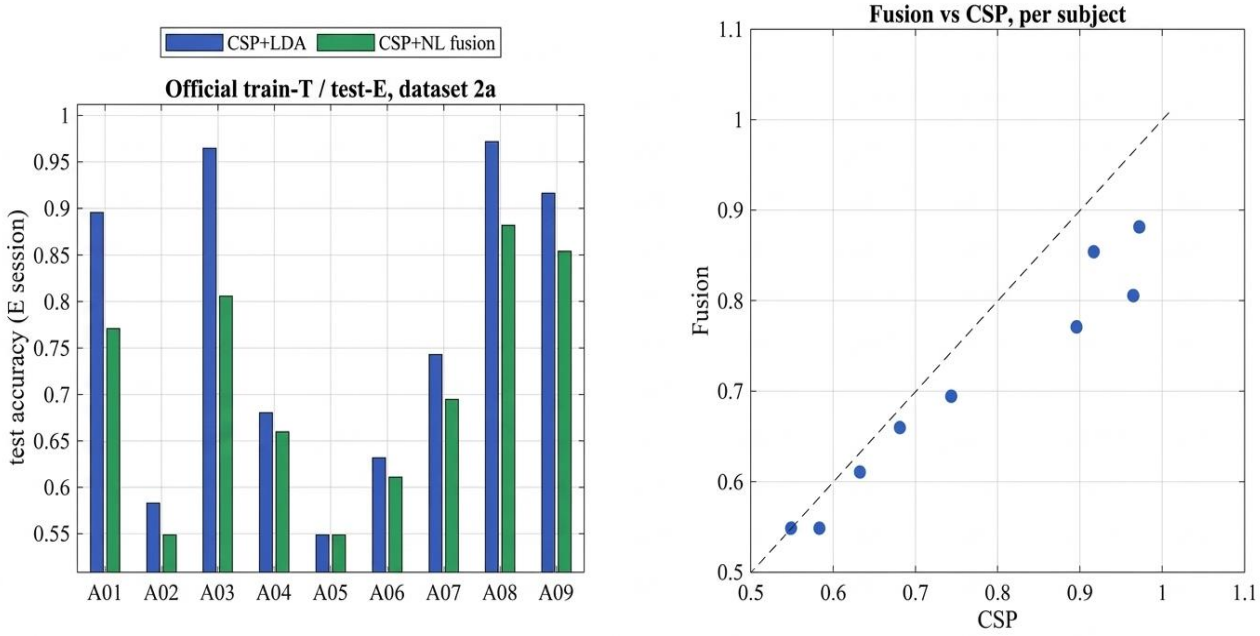


Fig. 9 Depicts official train-T/test-E evaluation, dataset 2a: per-subject test accuracy for CSP+LDA and the fusion (left), and the pooled two-class confusion matrix (right). Per-subject performance ranges from near chance to ~0.97, illustrating the large between-subject variability typical of these data.

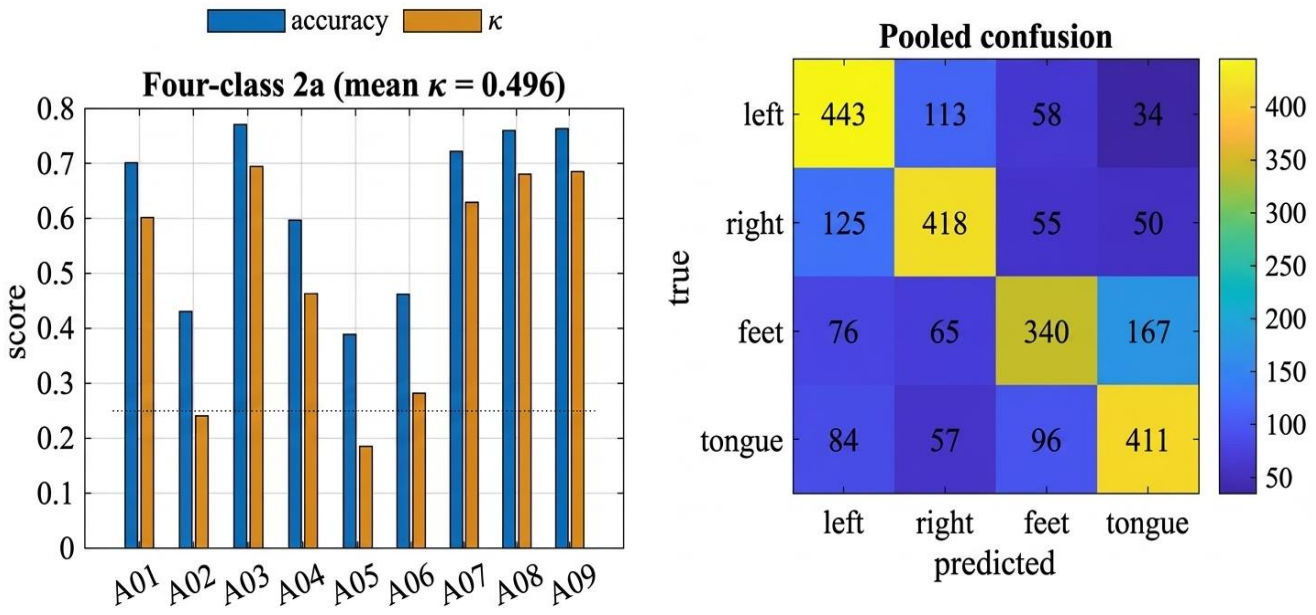


Fig. 10 Four-class 2a under the official protocol (one-versus-rest CSP + LDA). Per-Subject accuracy and Cohen's κ (left; dotted line = 0.25 chance) and the pooled 4 × 4 confusion matrix (right). Mean κ=0.496.

Table 4. Accuracy and Cohen's κ, dataset 2a, left vs. right, matched shrinkage-LDA, official train-T/test-E. d_z and q are paired tests against the band-power baseline across nine subjects, FDR-corrected across the table; band power itself is the reference and so has no entry in those two columns.

Feature Family	Val.	Test	κ	d_z vs. BP	q
Filter-bank CSP	0.809	**0.826**	0.653	+1.12	0.020
CSP + PAC	0.762	0.769	0.539	+1.23	0.020
CSP (spatial–spectral)	0.761	0.771	0.542	+1.02	0.024
Riemannian tangent space	0.740	0.749	0.499	+1.06	0.022
CSP + Nonlinear-7 (Set B)	0.712	0.718	0.435	+0.55	0.162
CSP + Nonlinear-7	0.701	0.708	0.417	+0.31	0.412

Band-power (spectral)	0.701	0.691	0.381	reference	reference
Bispectrum magnitude	0.705	0.701	0.401	+0.14	0.694
Nonlinear-7	0.596	0.576	0.153	−0.98	0.024
VMD-nonlinear	0.575	0.568	0.136	−0.98	0.024
Nonlinear-7 + PAC	0.593	0.569	0.137	−1.19	0.020
Nonlinear-7 (Set B)	0.577	0.563	0.125	−1.35	0.020
Bicoherence	0.541	0.524	0.048	−1.93	0.006
PAC-direct	0.522	0.535	0.069	−1.29	0.020

Table 5. Modern baselines, both datasets, official train-T/test-E, mean over nine subjects. *p* and *q* are paired tests against CSP.

Method	2a Acc. (sd)	2a κ	2b Acc. (sd)	2b κ
CSP + LDA	0.771 (0.169)	0.542	0.748 (0.138)	0.495
Filter-bank CSP (FBCSP)	0.826 (0.138)	0.653	0.786 (0.143)	0.571
Riemannian MDM	0.727 (0.180)	0.454	0.749 (0.144)	0.498
Riemannian tangent space + LDA	0.749 (0.148)	0.499	0.752 (0.136)	0.504
CSP + nonlinear fusion	0.708 (0.126)	0.417	0.759 (0.148)	0.517

Table 6. Surrogate-corrected residual decoding. "Raw" takes the descriptor values; "corrected" uses the per-trial surrogate *z*. *p* is a sign-flip test against chance; *q* is FDR-corrected across the four tests.

Dataset/descriptor set	Raw	Corrected	Band power	*p* (corr.)	*q*
2a / Set A	0.599	0.492	0.691	0.567	0.567
2a / Set B	0.579	0.553	0.691	0.005	0.021
2b / Set A	0.623	0.538	0.735	0.029	0.038
2b / Set B	0.621	0.561	0.735	0.023	0.038

Table 7. Per-descriptor IAAFT surrogate test, dataset 2a, 39 surrogates per trial. The first seven rows are Set A, the remainder Set B. A descriptor at $z \approx 0$ with ≈5% of trials individually significant carries nothing beyond the power spectrum.

Descriptor	Mean *z*	Trials sig. (%)	*p*	*q*
Higuchi FD	−0.09	6.5	0.092	0.138
Katz FD	−0.22	3.3	0.003	**0.011**
Petrosian FD	+0.04	5.7	0.376	0.470
DFA	−0.09	6.7	0.121	0.165
Sample entropy	−0.11	4.7	0.031	0.051
Permutation entropy	−0.01	5.8	0.872	0.872
Spectral entropy	−0.24	6.7	0.006	**0.017**
Time-reversal asymmetry	−0.29	6.4	0.027	0.051
Higuchi FD (k_{max}=8)	−0.29	9.2	0.0003	**0.004**
Prediction error	−0.28	7.5	0.001	**0.011**
Correlation dimension	−0.21	6.8	0.003	**0.011**
Lempel–Ziv complexity	+0.01	7.9	0.704	0.812

Table 8. Added value of the surrogate-corrected candidates over a filter-bank CSP + Riemannian ensemble, official protocol.

	Dataset 2a	Dataset 2b
Ensemble, mean test κ	0.6898	0.5861
Ensemble + candidates, mean test κ	0.6898	0.5782
Δκ (*p*)	+0.0000 (1.00)	−0.0079 (0.50)
Candidate share of the feature pool	33.3%	50.0%
Share of selected slots actually taken	**0.0%**	**4.2%**

The past work reported within-session validation results as uniformly optimistic. In these data, this is not the case. The sign of the difference depends on the dataset structure, so quoting a published within-session figure against a matched-protocol figure is invalid in both directions.

This paper also implemented the full four-class 2a problem under the official protocol with one-versus-rest CSP and LDA, to establish that the pipeline is competitive: mean accuracy 62.2%, Cohen's κ=0.496 against a chance level of 25%-in the region of the competition's winning FBCSP entry

(κ≈0.57) [16] for a plain spatial-filter baseline. Subsequent negative findings, therefore, cannot be attributed to a weak pipeline (see Figures 8–10).

4.5. Isolated Decoding: Every Nonlinear Family Falls below Band Power

Table 4 reports every family in dataset 2a under the official protocol, with each family tested against the band-power baseline and the whole set corrected by FDR. Every nonlinear and higher-order family lands below band power, and for all of them, the gap is significant after correction.

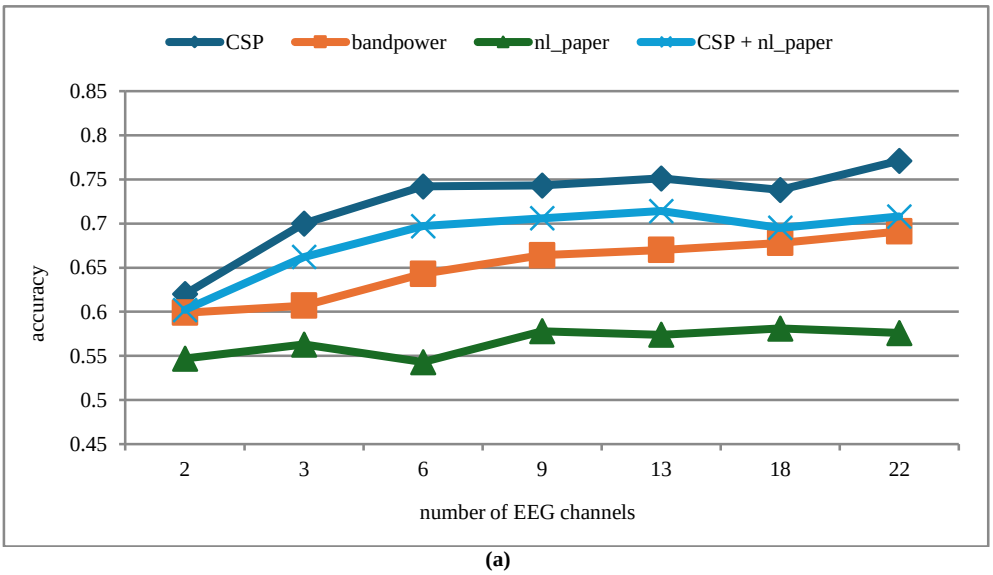
The best nonlinear family reaches 0.576 against 0.691 for the simplest spectral feature, let alone 0.771 for CSP or 0.826 for FBCSP. On 2b, the ordering is the same, though the margins are narrower-Nonlinear-7 at 0.668 against band power at 0.735, a gap that does not reach significance after correction (q=0.123)-and Section 4.8 takes up why the two datasets differ. This point can be clarified in three ways. In the first place, each family's validation and test tracks are carefully watched, which means that the ordering is not due to a cross-validation artefact. Secondly, the outcome is not limited to the seven descriptors mentioned in Section 3.2: Set B, in fact, scores slightly lower, at 0.563. Thirdly, bispectrum magnitude is the only higher-order family that manages to decode, doing so at a value of 0.701-this being statistically no different from band power-and Section 4.10 gives the exact reason for this. Fusion adds nothing. CSP combined with Nonlinear-7 scores 0.708 as compared to CSP's 0.771, and this decline is significant (in the modern-baseline comparison, p=0.008 and q=0.031). The addition of 154 complexity features to the six existing CSP features results in a loss of six points-a pattern characteristic of redundant and noisy factors that increase variance. On the other hand, CSP combined with PAC cannot be distinguished from CSP (0.769 versus 0.771) since the classifier just ignores the four PAC dimensions (Table 5).

4.6. Does Anything Survive the Spectral Control?

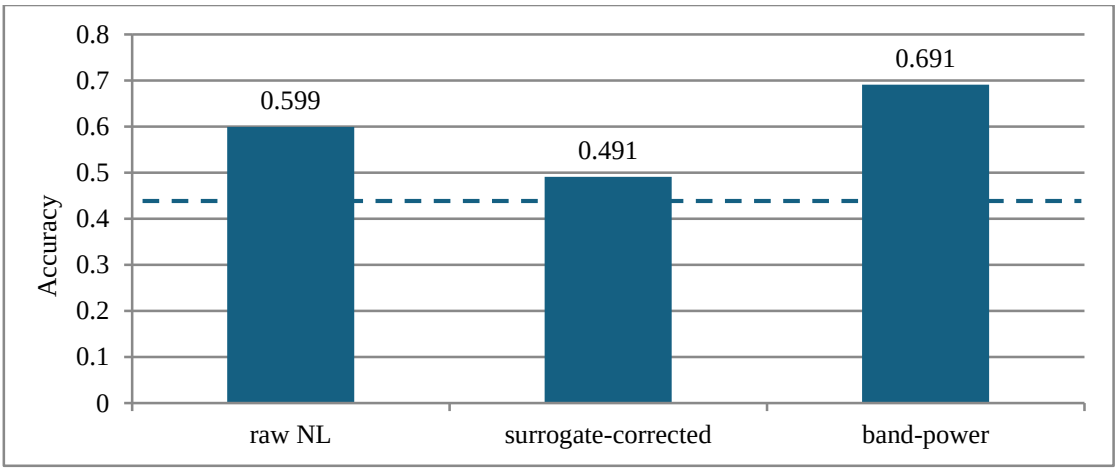
The decoding tables show that nonlinear features are worse. They do not show why. For that, each descriptor is replaced by its surrogate-corrected residual-the standardized z-score of the measured feature relative to the IAAFT surrogates generated from the same trial.

Table 9. Indicates Cohen's κ for dataset 2a, according to the official protocol. The electrode-matched pairs (mono4/bip2 and mono6/bip3) serve to separate the derivation from the number of sensors. n/r means that the four-class condition was analysed using one-versus-rest CSP combined with LDA alone, since the issue in that case is whether a low-channel montage can at all represent both the feet and the tongue, not which feature family is better; the two-class section above answers that question for each family.

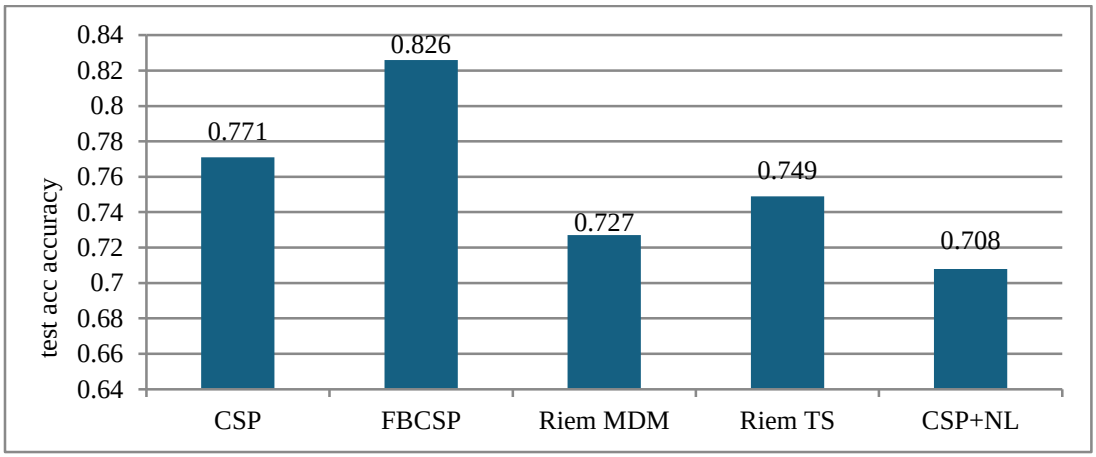
Montage	Ch.	Elec.	CSP	Band power	Nonlinear-7
Two-class					
mono22	22	22	0.542	0.381	0.153
bip3	3	6	0.414	0.420	0.290
mono4	4	4	0.407	0.244	0.168
bip2	2	4	0.346	0.418	0.273
Four-class (one-vs-rest CSP + LDA, chance κ=0)					
mono22	22	22	0.496	n/r	n/r
mono6	6	6	0.364	n/r	n/r
bip3	3	6	0.275	n/r	n/r



(a)



(b)



(c)

Fig. 11 Controlled re-evaluation of BCI IV-2a. (a) Channel-density sweep: nonlinear features never reach CSP or band power at any channel count, (b) Surrogate-controlled test, and (c) Modern baselines under the official protocol: filter-bank CSP leads; adding nonlinear features to CSP loses.

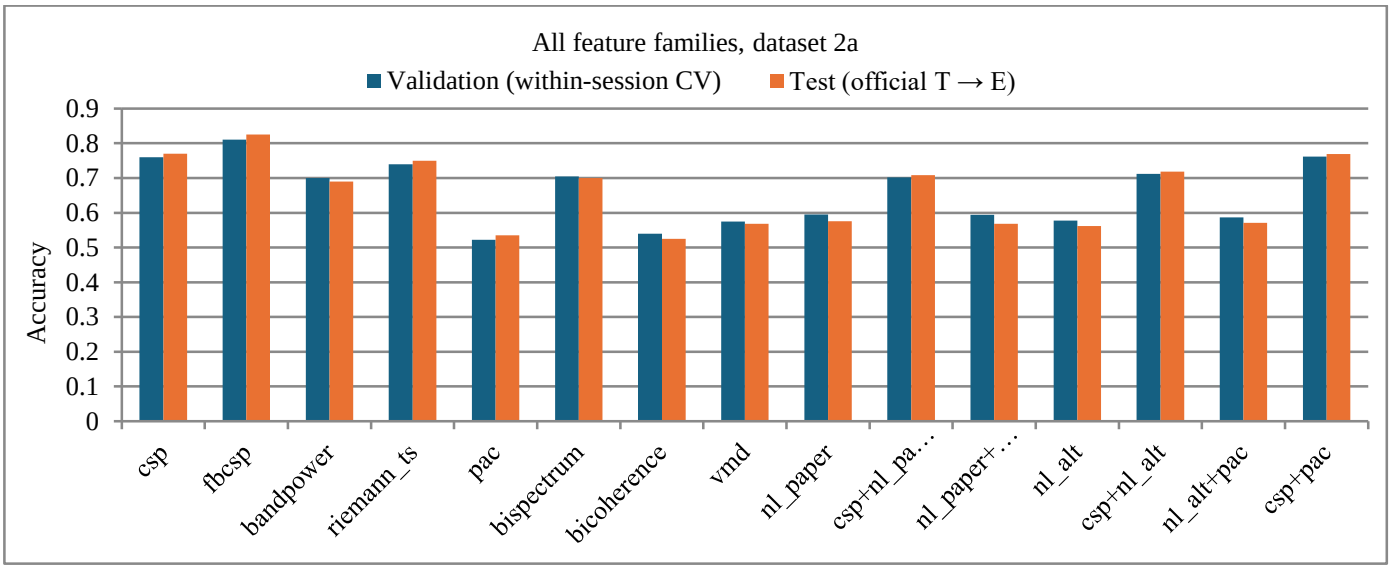


Fig. 12 Left-versus-right decoding on BCI IV-2a, all feature families, within-session validation (blue) versus official independent-session test (orange). Spectral and spatial baselines dominate; nonlinear and higher-order families fall at or below band power; fusion adds nothing. Validation and test track closely; therefore, the ordering is not a cross-validation artifact.

What remains is only what the trial's own power spectrum cannot explain, and decode that (Table 6). A partial positive result that runs counter to the study's own hypothesis was found and is reported. In three of the four dataset × descriptor-set combinations, the corrected residual decodes above chance, and it survives FDR correction. A strict claim that nonlinear descriptors contain nothing beyond the power spectrum is therefore not supportable, and it is not made here. The surviving residual clears the chance by 3.8 to 6.1 percentage points; band power clears it by 19 to 24 percentage points. Whatever non-spectral structure these descriptors detect is real and roughly a quarter the magnitude of the plainest spectral feature-and, as the next subsection shows, it is worth precisely nothing once a competent pipeline is already in place.

Consistent with this, Table 7 applies the surrogate test to each descriptor individually. Several depart significantly from their spectrally matched null after correction-Katz FD, spectral entropy, Higuchi with kmax=8, prediction error and correlation dimension on 2a, and most descriptors on 2b-but every significant departure is negative, meaning the real trials are less complex than linear surrogates, and the fraction of individually significant trials sits at 3–9%, i.e. at the nominal 5% rate. The descriptors are detecting something, not a rich nonlinear structure waiting to be decoded.

4.7. The Decisive Test: Added Value over a State-of-the-Art Pipeline

The real question a practitioner has to deal with is not whether a feature contains information, but whether adding it to an existing system is of any benefit. For this reason, a deliberately strong baseline was constructed-namely, a filter-bank CSP with Euclidean Alignment, combined with a filter-bank Riemannian tangent space, the latter being weighted by cross-validated κ-and then put the surrogate-corrected nonlinear and PAC candidates into the same mutual-information feature pool so that they had to compete with each other on the basis of their merit (Table 8). On dataset 2a, the fused and backbone columns are bit-identical for all nine subjects: with mutual-information selection, given a pool of one-third candidates, not a single one was admitted. The added value is not small; it is exactly zero. The redundancy is defined by this combination: although the candidates are individually related to the label-18 passing the per-feature label-permutation gate in 2a ($p=0.0013$) and 25 in 2b ($p=5.8\times10-7$), many more than the 5% to be expected by chance-none of them is ever selected. The fact that they are reliably label-related yet are never chosen constitutes redundancy; whatever knowledge these features have about the label is already known by the spectral features.

4.8. Channel Count versus Channel Derivation

The main argument in favour of complex features is that they are justified in cases where there are few channels, and the typical evidence shows that they appear more competitive on dataset 2b than on dataset 2a when there are three channels rather than twenty-two. However, this comparison conflates three factors: the number of electrodes (22 versus 6), the type of spatial derivation (monopolar versus bipolar), and the dataset itself (which involves different subjects, sessions, hardware, and feedback runs). To address this, they are treated separately. No low-channel crossover. Sweeping nested monopolar montages from 2 to 22 channels, the ordering CSP > band power > nonlinear holds at every channel count from 3 upwards, and nonlinear features are last at all seven densities (2 ch: 0.620 / 0.600 / 0.547; 22 ch: 0.771 / 0.691 / 0.576). At the two-channel point, band power and CSP are within 2 points of each other, so a strict ordering is not claimed there; the claim that is made-that complexity features never become competitive-holds throughout.

The derivation is not the explanation. Comparing mono4 with bip2-the same four electrodes, differenced or not-the bipolar derivation is worth +0.174 κ to band power and only +0.105 κ to Nonlinear-7. Spatial focusing helps the spectral feature more than the complexity feature, which is the opposite of what the low-channel argument requires. With the montage matched, the dataset gap remains. At bip2-two channels and four electrodes on both datasets-2b still exceeds 2a on every family (band power κ 0.465 vs 0.418; CSP 0.469 vs 0.346; Nonlinear-7 0.323 vs 0.273). What separates the datasets is therefore not spatial focusing but electrode spacing (≈2.5 cm as recorded on 2b against roughly twice that for a derived FC−CP pair), subjects, sessions, and 2b's feedback runs. Nonlinear accuracy achieves band-power accuracy across the ladder. Overall montages, the per-montage κ of Nonlinear-7 correlates with that of band power at $r=0.77$ on 2a and $r=0.84$ on 2b. Change the montage and the two families move together-which is what "spectral proxy" means, stated without any surrogate machinery at all (Table 9).

The low-channel montage is task-specific, not cheap: With regard to the two-class problem, the transition from mono22 to bip3 yields a CSP cost of 0.128κ. In the case of the four-class problem, the same reduction leads to a cost of 0.221-yet the electrode-matched control breaks this down by showing that going from 22 to 6 electrodes causes a cost of 0.132 ($p=0.007$), and the bipolar derivation adds a further 0.089 ($p=0.0005$, $dz=-1.90$). The left-versus-right comparison favours a central strip, since C3/C4 is nearly the optimal montage for that particular contrast; the feet activate Cz, and the tongue shows diffuse activation; three central channels are not sufficient to represent these. Per-class recall at bip3 drops from 0.52–0.68 to 0.43–0.51 compared to mono22. Hence, any assertion that a small number of channels is sufficient must be made for a specific task and not in general. Cz is a working negative control: Dropping Cz from the bipolar montage (bip3 → bip2) changes Nonlinear-7 by −0.017 κ on 2a and −0.013 on 2b-indistinguishable from zero, exactly as a channel carrying no left/right information should behave (see Figures 11 and 12).

Table 10. Higher-order spectra on the C3/C4 pair, pooled over all 18 subjects of both datasets. p-values are paired across subjects and FDR-corrected within the table

Quantity	Accuracy	Surrogate *z*	Verdict
Band power (baseline)	0.721	n/a ‡	reference
Bispectrum magnitude	0.715	+25.5 †	indistinguishable from band power, p=0.59
Bicoherence	0.546	−0.16	*below* band power, $p=3.5\times10^{-6}$; below its own null

† The magnitude *z* is confounded and should not be read as evidence of nonlinearity: the bispectrum is averaged over overlapping 128-sample segments while IAAFT is applied to the whole epoch, so segment-level spectra are not preserved, and the surrogate's triple products add incoherently. The statistic measures within-epoch non-stationarity. Bicoherence divides this scaling out and is the interpretable quantity.

‡ Band power has no surrogate *z*, and this is a property of the test rather than a missing measurement. IAAFT preserves each trial's amplitude spectrum *exactly*, and band power is a function of that spectrum; therefore, $Q(\text{data})=\langle Q(\text{surrogate})\rangle$ identically and both the numerator and the denominator of *z* vanish. Band power is the quantity the other two rows are tested *against*, which is also why it carries no verdict of its own.

4.9. Nonlinearity does not Predict Benefit

Tying the two halves of the study together, a subject's measured EEG nonlinearity-the surrogate z of the time-reversal statistic from Section 4.2-does not predict how much that subject gains from nonlinear features: Spearman ρ=−0.23, p=0.56 across the pooled 18 subjects. Even subjects whose EEG rejects the linear null gain nothing extra. Given the instability of the nonlinearity estimate documented in Section 4.2, this is regarded as a supporting observation rather than a load-bearing one.

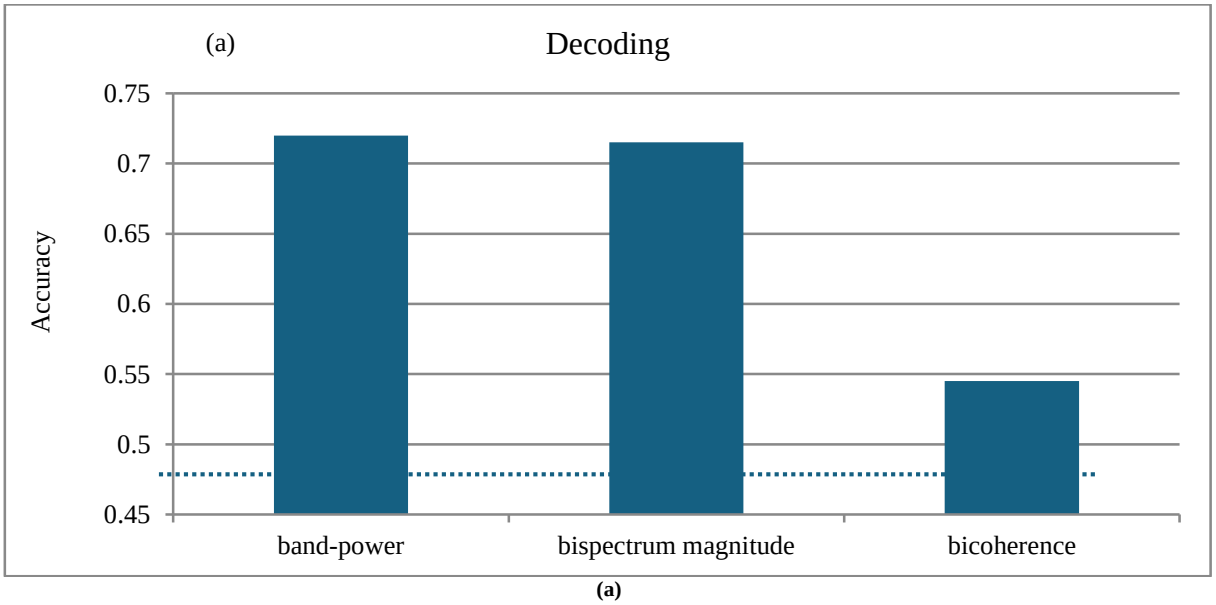
4.10. The Mechanism, Made Explicit: Bispectrum versus Bicoherence

A higher-order spectrum decomposes into two parts that separate the question cleanly. The raw bispectrum magnitude |B(f1,f2)| scales with the power at its constituent frequencies; the normalized bicoherence b2(f1,f2) divides that scaling out and is zero for any linear Gaussian process. If the decodable content of higher-order spectra is really power in disguise, magnitude should decode, and bicoherence should not. That is exactly what happens, and here it can be stated across all 18 subjects, rather than the single-subject analysis numbers make the main point:

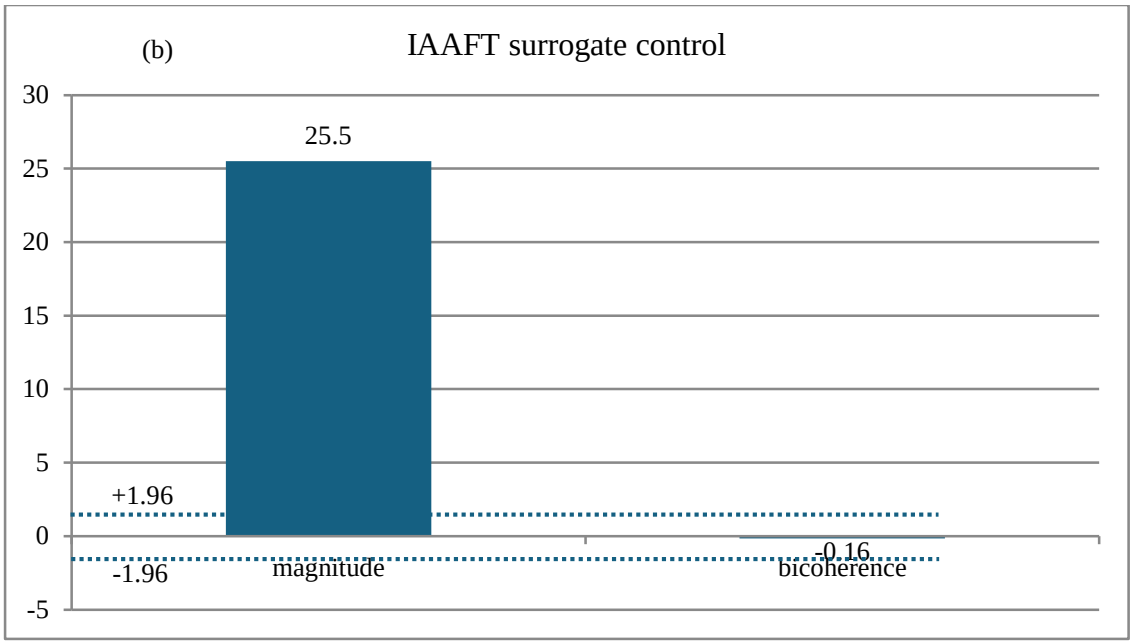
Bispectrum magnitude: 71.5% accuracy, compared with 72.1% for band power. The difference is not statistically significant (p = 0.59), so they perform essentially the same. Across subjects, the two measures are very strongly correlated (r = 0.958). In practical terms, this means the bispectrum magnitude captures almost the same information as band power.

Bicoherence: When the power-related scaling is removed, performance drops to 54.6%, which is significantly worse than band power ($p = 3.5 \times 10^{-6}$). It also performs no better than its spectrally matched null (z = −0.16), and only 6.3% of trials are individually significant-about what you would expect by chance.

The apparent decoding advantage of bispectrum magnitude seems to come from the power information already contained in it, not from genuinely new information captured by phase relationships. Remove the power, and nothing remains. This is the clearest statement in the study of what the complexity and higher-order families are actually measuring, and it does not depend on any claim about fractal dimensions or entropies (see Figure 13).



(a)



(b)

Fig. 13 Higher-order spectra of the bipolar C3/C4 pair. (Left) Decoding accuracy-bispectrum magnitude tracks band power, while bicoherence falls to near-chance levels. (Right) IAAFT surrogate *z* for bicoherence, which does not exceed the spectrally matched null. Magnitude decodes; once its power scaling is divided out, nothing remains.

4.11. Robustness to Ocular Correction

Since complexity measures are based on broadband sensorimotor signals, they are most affected by ocular contamination-an artefact that might cause their apparent values to be overestimated. When the comparison was redone with the three given EOG channels removed from the continuous EEG prior to trial extraction, both the order and all the conclusions remained unchanged. Regarding Figure 2a: CSP changed from 0.759 to 0.749, band power from 0.674 to 0.668, and nonlinear features from 0.602 to 0.593, with no change being significant (in all cases q=0.65). As for Figure 2b: CSP went from 0.715 to 0.720, band power from 0.714 to 0.708, and nonlinear features from 0.638 to 0.636 (in all cases $q \geq 0.71$).

In fact, the EOG correction tends to reduce the nonlinear components and place them at the end, which is consistent with the idea that a portion of their variance is attributable to non-cortical sources rather than representing a genuine nonlinear signal. This thus supports the negative conclusion instead of refuting it.

5. Discussion

In a study involving fifteen feature families (with 18 subjects), and using two public benchmarks, five montages, two task settings and a matched classifier according to the official 'independent session' protocol, it was found that nonlinear and higher-order features contain no information beyond that given by the power spectrum; individually they are at a level below that of simple band power, lose six points when combined with CSP, are never chosen when compared against a state-of-the-art feature pool, show no evidence of low-channel crossover, and offer no benefit to the subjects whose EEG rejected a linear null.

What was found does not fit the simplest version of this claim: A surrogate-corrected residual decodes above chance in three of four dataset × descriptor-set combinations and survives FDR correction (Section 4.6). This is reported rather than rounded away. Nonlinear descriptors are therefore not *pure* spectral proxies-something in them is not explained by the trial's own power spectrum. But that something clears chance by 4–6 points, where band power clears it by 19–24, and it adds exactly 0.0000 κ to a competent pipeline while being selected into 0.0% of available slots. The accurate way of putting it is not "there is nothing there" but "there is nothing there worth using", and it is important to make this distinction since the first statement can be refuted by a single decisive test, whereas the second is what a practical person actually needs to know.

The most convincing evidence is to be found in the selection results rather than in the accuracy tables. Individually, each of the candidate features is reliably associated with the labels: 18 of them pass the per-feature label permutation test on 2a (p=0.0013) and 25 on 2b. Yet the mutual-information selection method results in a subset which is one-third the size of the original pool, rejects all the features, and produces fused and backbone outcomes that are identical bit for bit for each subject. It is precisely what redundancy entails that features should be related to the labels yet never be selected, and this is a statement which has nothing to do with

the classifier: it is mutual information, not the discriminant that is the criterion used for selection.

The mechanism can be seen in one decomposable quantity. The magnitude of the bispectrum performs just as well as the band power (0.715 compared with 0.721, p=0.59), and the accuracy obtained for each subject with respect to the bispectrum is correlated with the accuracy obtained with respect to band power (r=0.958). If the power scaling is removed, the bicoherence drops below both the band power and its own surrogate null. To determine what these families measure, one need not consider fractal dimensions.

Comparison with the state of the art: The pipelines used here perform at the level expected of the current state of the art rather than below it, and this is what gives the negative result its force. It is reported on the four-class 2a problem, the one-versus-rest CSP baseline reaches κ=0.496, in the region of the competition-winning FBCSP entry ($\kappa\approx$0.57) [16]; on the two-class problem, the FBCSP implementation reaches 0.826 accuracy (κ=0.653) on 2a and 0.786 (κ=0.571) on 2b, meeting or exceeding the CSP, Riemannian MDM, and tangent-space baselines of Table 5. The accuracy figures cited by the nonlinear studies in Table 1 (about 78–92%) are achieved only when using within-session, best-subject, or fusion reporting; but when the same feature families are re-implemented under the official independent-session protocol with a classifier that matches the one used before, the accuracy drops to 0.52–0.71 (as shown in Table 4). The improvement achieved in the current evaluation over the previous state of the art is therefore methodological rather than numerical in nature, since it reaches state-of-the-art decoding levels using the spectral and spatial baselines while also incorporating the surrogate, selection, and montage controls, which show where the previously reported gains come from. Studies that claim to have achieved a nonlinear improvement can be directly compared with this framework, as each of the checks is available as a one-line addition to the pipeline.

It does not go against the published accuracy figures; instead, it offers a different interpretation of them. The successes reported for the bispectrum, PAC, and fractal dimension in Table 1 correspond to the accuracy rates of tuned, fused, all-channel, within-session systems in which the nonlinear feature is propagated along a spatial–spectral pipeline that performs the discriminative function. However, when these results are assessed individually on the sensorimotor channels where the effect is said to occur, using a split free from leakage, together with a surrogate control and applying a multiple-comparison correction, the apparent advantage vanishes.

There are two assertions in the literature on reassessment that the present data do not support. The first is that within-session validation is generally optimistic; on both datasets, the results from the official independent-session test outperformed those from within-session cross-validation, the reason being structural on dataset 2b and due to noise on dataset 2a.

The second claim that the competitiveness of complexity features on low-channel montages is due to spatial focusing, is not correct: the bipolar derivation has a greater effect on band power than on complexity, and even when the montage is fully matched, the gap between the datasets remains. Just as the work they are reassessing can adopt a convenient explanation, so too can the reassessments.

On the dynamical framing: One dynamical finding is retained-no correlation-dimension saturation in any of the 18 subjects-and the rest is explicitly not built upon. Delay, embedding dimension, and the count of subjects rejecting the linear null all vary substantially under an arbitrary channel derivation (Section 4.2). Reporting a specific embedding dimension or a specific count of “nonlinear subjects” from 2-second scalp epochs would repeat, in the opposite direction, the error this paper is about.

Testing discipline: Before a nonlinear MI feature is credited with an improvement, it should: (i) be evaluated against a spectral baseline of comparable dimensionality; (ii) be reported in isolation, not only in fusion; (iii) be tested on an independent session rather than within-session folds; (iv) be corrected for the number of descriptors actually tested; (v) be offered into the feature pool of a competent baseline rather than concatenated onto a weak one; and (vi) pass an IAAFT surrogate redundancy test with the surrogate generated *before* any decimation and with enough surrogates for the rank test to reach α.

The first three are widely acknowledged in principle and widely omitted in practice. The last three are, to the authors’ knowledge, not standard at all in this literature, and (v) is the one that separates “contains information” from “adds value”. The full pipeline is released so that each check is a one-line addition to any feature-evaluation study.

6. Limitations

The study tested only a specific left/right Motor-Imagery (MI) task and a four-class control task, focusing on the channels where the relevant brain rhythm was strongest. So, the results do not prove that nonlinear information is useless in every situation. Nonlinear features might still be helpful in other types of brain-imagery tasks, at different frequency ranges, or when the EEG data are reconstructed to estimate activity from specific brain sources (source space).

The classifier is kept fixed, and that has a double effect. All the families are analysed using the same shrinkage LDA. This eliminates the most common confounding factor in the literature on nonlinear features-attributing greater performance to a feature-, but it does mean that the decoding

statements are, strictly speaking, concerned only with information that can be used by a linear discriminant.

If a compact convolutional network were applied to the raw signals or to the phase-space or recurrence images, the objection that 'perhaps deep learning detects the nonlinearity' would be overcome, and this would represent the natural progression. It should be pointed out that the selection result given in Section 4.7 is not affected by this limitation, since mutual-information ranking makes no assumption about linearity.

The linear-stochastic null hypothesis is rejected by the surrogate test, and the z-score is based on a surrogate distribution that is approximately Gaussian; for strictly non-negative measures, the authors therefore also report the rank-based fraction of significant trials, since this method makes no such assumption. The estimates of the correlation dimension are calculated using segments much shorter than the conventional recommendation for absolute dimension estimation; they are used only as saturation tests, not as estimates of attractor dimension, and the finding that the system is not saturated is robust to this because it relates to the shape of the D2 versus m plot rather than to the value of that shape.

The remaining result is at the limit of the study's resolution. Three of the four residual tests pass the FDR criterion, but the effects are 4 to 6 points, with 9 observations per dataset. It would not be surprising if the exact pattern of which combinations pass changed on new data; it would, however, be surprising if the magnitude of the effects changed. The electrode spacing for the 2a bipolar pair is about twice that of the 2b recorded pairs, which means that the 'matched montage' referred to in Section 4.8 is one that has the same number of channels, the same number of electrodes, and the same derivation type, not one that is matched in spatial scale. The pooled 2a+2b statistics, therefore, include two spatial transfer functions.

Lastly, the statistical power at the group level is restricted to nine subjects per dataset. A small genuine effect could have been missed. The statement being made is that there is "no usable improvement under fair evaluation", not that "no effect exists".

7. Conclusion

By applying chaos- and fractal-based analysis, the EEG is regarded as an observable of a nonlinear dynamical system, with fractal dimensions, correlation dimension, Lyapunov exponents, and entropy measures being added to the traditional spectrum. The original goal of the field-namely, low-dimensional deterministic brain chaos-has been replaced by a more cautious perspective, and the present data confirm this: in none of the 18 subjects does the correlation dimension reach saturation. The analysis does not go any further along dynamical lines since the remaining phase-space estimates vary too much when an arbitrary channel derivation is used to support any conclusion.

With regard to motor-imagery decoding, the conclusion is clear. In tests carried out under surrogate control using independent sessions and a matched classifier with multiple-comparison correction, nonlinear and higher-order features contain no useful information beyond that provided by the power spectrum. They are no better than simple band power; when combined with CSP, they lose six points; none of the available slots in a state-of-the-art feature pool selects them; they do not become competitive when the number of channels is reduced; and they provide no help even in those subjects for whom the EEG rejects a linear null. Although a small non-spectral residual is statistically detectable, it contributes precisely 0.0000 κ; the finding is that this combination, not the absence of any effect, is the result.

Their function is to provide a description-that is, to characterise the sensorimotor dynamics and their subject-dependent nonlinearity-rather than to make predictions. The reusable IAAFT redundancy test offered here, combined with the rule that a candidate feature must compete within the selection pool of a competent baseline, turns this observation into a practical criterion: whenever a feature is claimed to capture nonlinearity, it must pass both tests before a nonlinear gain is awarded.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

This research received no specific grant from any funding agency.

Data Availability Statement

The BCI Competition IV datasets 2a and 2b analyzed in this study are publicly available from the Graz University of Technology at https://www.bbci.de/competition/iv/. Every number reported in this paper is produced by a single self-contained MATLAB program from the raw GDF files.

References

[1] G. Pfurtscheller, and F.H. Lopes da Silva, "Event-Related EEG/MEG Synchronization and Desynchronization: Basic Principles," *Clinical Neurophysiology*, vol. 110, no. 11, pp. 1842-1857, 1999. [CrossRef] [Google Scholar] [Publisher Link]

[2] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal Spatial Filtering of Single Trial EEG during Imagined Hand Movement," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 4, pp. 441-446, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[3] A. Babloyantz, J.M. Salazar, and C. Nicolis, "Evidence of Chaotic Dynamics of Brain Activity during the Sleep Cycle," *Physics Letters A*, vol. 111, no. 3, pp. 152-156, 1985. [CrossRef] [Google Scholar] [Publisher Link]

[4] Edward N. Lorenz, "Deterministic Nonperiodic Flow," *Journal of the Atmospheric Sciences*, vol. 20, no. 2, pp. 130-141, 1963. [CrossRef] [Google Scholar] [Publisher Link]

[5] B.B. Mandelbrot, *The Fractal Geometry of Nature/Revised and Enlarged Edition/*, W.H. Freeman and Co, New York, 1983. [Google Scholar] [Publisher Link]

[6] James Theiler et al., "Testing for Nonlinearity in Time Series: The Method of Surrogate Data," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1-4, pp. 77-94, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[7] Thomas Schreiber, and Andreas Schmitz, "Improved Surrogate Data for Nonlinearity Tests," *Physical Review Letters*, vol. 77, no. 4, pp. 635-638, 1996. [CrossRef] [Google Scholar] [Publisher Link]

[8] Klaus Lehnertz, and Christian E. Elger, "Can Epileptic Seizures be Predicted? Evidence from Nonlinear Time Series Analysis of Brain Electrical Activity," *Physical Review Letters*, vol. 80, no. 22, pp. 5019-5022, 1998. [CrossRef] [Google Scholar] [Publisher Link]

[9] Klaus Lehnertz, and Christian E. Elger, "Spatio-Temporal Dynamics of the Primary Epileptogenic Area in Temporal Lobe Epilepsy," *Electroencephalography and Clinical Neurophysiology*, vol. 95, no. 2, pp. 108-117, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ralph G. Andrzejak et al., "Indications of Nonlinear Deterministic and Finite-Dimensional Structures in Time Series of Brain Electrical Activity," *Physical Review E*, vol. 64, no. 6, pp. 1-8, 2001. [CrossRef] [Google Scholar] [Publisher Link]

[11] Denis Jordan et al., "Electroencephalographic Order Pattern Analysis for the Separation of Consciousness and Unconsciousness," *Anesthesiology*, vol. 109, no. 6, pp. 1014-1022, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[12] Amr F. Mohamed, and Vacius Jusas, "Advancing Fractal Dimension Techniques to Enhance Motor Imagery Tasks using EEG for Brain–Computer Interface Applications," *Applied Sciences*, vol. 15, no. 11, pp. 1-27, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[13] Venkatesh Kanagaluru, and M. Sasikala, "Two Class Motor Imagery EEG Signal Classification for BCI using LDA and SVM," *Signal Processing*, vol. 41, no. 5, pp. 2743-2749, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[14] Chengzhen Wu et al., "The Application of Entropy in Motor Imagery Paradigms of Brain–Computer Interfaces," *Brain Sciences*, vol. 15, no. 2, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[15] Alois Schlögl et al., "A Fully Automated Correction Method of EOG Artifacts in EEG Recordings," *Clinical Neurophysiology*, vol. 118, no. 1, pp. 98-104, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[16] Kai Keng Ang et al., "Filter Bank Common Spatial Pattern Algorithm on BCI Competition IV Datasets 2a and 2b," *Frontiers in Neuroscience*, vol. 6, pp. 1-9, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[17] Peter Grassberger, and Itamar Procaccia, "Measuring the Strangeness of Strange Attractors," *Physica D: Nonlinear Phenomena*, vol. 9, no. 1-2, pp. 189-208, 1983. [CrossRef] [Google Scholar] [Publisher Link]

[18] Michael T. Rosenstein, James J. Collins, and Carlo J. De Luca, "A Practical Method for Calculating Largest Lyapunov Exponents from Small Data Sets," *Physica D: Nonlinear Phenomena*, vol. 65, no. 1-2, pp. 117-134, 1993. [CrossRef] [Google Scholar] [Publisher Link]

[19] Floris Takens, "Detecting Strange Attractors in Turbulence," *Dynamical Systems and Turbulence Warwick 1980*, Proceedings of a Symposium Held at the University of Warwick 1979/80, Heidelberg, Berlin, Springer, vol. 898, pp. 366-381, 1981. [CrossRef] [Google Scholar] [Publisher Link]

[20] Tomoyuki Higuchi, "Approach to an Irregular Time Series on the Basis of the Fractal Theory," *Physica D: Nonlinear Phenomena*, vol. 31, no. 2, pp. 277-283, 1988. [CrossRef] [Google Scholar] [Publisher Link]

[21] Joshua S. Richman, and J. Randall Moorman, "Physiological Time-Series Analysis using Approximate Entropy and Sample Entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039-H2049, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[22] Christoph Bandt, and Bernd Pompe, "Permutation Entropy: A Natural Complexity Measure for Time Series," *Physical Review Letters*, vol. 88, no. 17, pp. 1-5, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[23] C.-K. Peng et al., "Quantification of Scaling Exponents and Crossover Phenomena in Nonstationary Heartbeat Time Series," *Chaos*, vol. 5, no. 1, pp. 82-87, 1995. [CrossRef] [Google Scholar] [Publisher Link]

[24] Adriano B.L. Tort et al., "Measuring Phase-Amplitude Coupling between Neuronal Oscillations of Different Frequencies," *Journal of Neurophysiology*, vol. 104, no. 2, pp. 1195-1210, 2010. [CrossRef] [Google Scholar] [Publisher Link]

[25] C.L. Nikias, and M.R. Raghuveer, "Bispectrum Estimation: A Digital Signal Processing Framework," *Proceedings of the IEEE*, vol. 75, no. 7, pp. 869-891, 1987. [CrossRef] [Google Scholar] [Publisher Link]

[26] Alexandre Barachant et al., "Multiclass Brain-Computer Interface Classification by Riemannian Geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920-928, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[27] Konstantin Dragomiretskiy, and Dominique Zosso, "Variational Mode Decomposition," *IEEE Transactions on Signal Processing*, vol. 62, no. 3, pp. 531-544, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[28] J. Richard Landis, and Gary G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159-174, 1977. [CrossRef] [Google Scholar] [Publisher Link]

[29] Benjamin Blankertz et al., "Single-Trial Analysis and Classification of ERP Components—A Tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814-825, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[30] He He, and Dongrui Wu, "Transfer Learning for Brain-Computer Interfaces: A Euclidean Space Data Alignment Approach," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 2, pp. 399-410, 2020. [CrossRef] [Google Scholar] [Publisher Link]