

Article

Not peer-reviewed version

From Segment-Level Discrimination to Unseen-Subject Transfer in Motor Imagery EEG: A Subject-Wise Study of a GAF-PLV Parallel CNN

Renjie Lv , [Hesam Akbari](#) , Muhammad Tariq Sadiq ^{*} , [Wenwen Chang](#) ^{*} , Rab Nawaz

Posted Date: 28 April 2026

doi: 10.20944/preprints202604.1922.v1

Keywords: motor imagery EEG; cross-subject generalization; subject-independent evaluation; leave-one-subject-out cross-validation; Gramian Angular Field; phase-locking value; convolutional neural network



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

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.

Article

From Segment-Level Discrimination to Unseen-Subject Transfer in Motor Imagery EEG: A Subject-Wise Study of a GAF–PLV Parallel CNN

Renjie Lv ^{1,†}, Hesam Akbari ^{2,†}, Muhammad Tariq Sadiq ^{3,*}, Wenwen Chang ^{1,*} and Rab Nawaz ³

- ¹ School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
- ² Department of Information Science, University of North Texas, Denton, TX 76207, USA
- ³ School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, United Kingdom
- * Correspondence: m.t.sadiq@essex.ac.uk (M.T.S.); changww2013@126.com (W.C.)
- [†] These authors contributed equally to this work.

Abstract

Deep learning for motor imagery electroencephalography (MI-EEG) has repeatedly reported near-ceiling accuracy on public benchmarks. However, many studies use data partitioning strategies in which windows, images, or pooled samples from the same subject can influence both model development and evaluation, so the resulting numbers do not answer the stronger question of cross-subject generalization. This study presents a subject-wise leave-one-subject-out cross-validation (LOSOCV) analysis of a Gramian Angular Field (GAF) and Phase-Locking Value (PLV) parallel convolutional neural network developed for MI-EEG representation learning. Under segment-level splitting, the framework achieved 99.73% binary accuracy in proof-of-concept benchmarking. Here, the same feature-construction logic is examined under LOSOCV on the 105 retained PhysioNet subjects for which subject-wise rerun outputs were available in this study. Under this subject-wise setting, the model achieves $58.07\% \pm 8.27\%$ mean accuracy, $53.48\% \pm 11.19\%$ macro-F1, and 0.1615 ± 0.1654 Cohen's kappa, with held-out subject accuracy ranging from 38.10% to 78.57%. Relative to the earlier segment-wise benchmark, the mean generalization gap is 41.66 points, while held-out-subject gaps span 21.16–61.63 points. By combining complete held-out-fold disclosure, retained-cohort accounting, bootstrap confidence intervals, and explicit protocol-sensitive comparison, the study provides a stronger subject-wise reference point for future MI-EEG evaluation and a more defensible basis for interpreting translational claims. The significance of the work lies in showing how a strong proof-of-concept benchmark behaves under a subject-wise inference target and in providing a clearer field reference for subject-independent MI-EEG research.

Keywords: motor imagery EEG; cross-subject generalization; subject-independent evaluation; leave-one-subject-out cross-validation; Gramian Angular Field; phase-locking value; convolutional neural network

1. Introduction

Electroencephalography (EEG) remains one of the most practical modalities for brain–computer interface (BCI) research because it is non-invasive, portable, and temporally precise [1,2]. Within EEG-based BCI, motor imagery (MI) is particularly important because it supports assistive control, neurorehabilitation, and human–machine interaction without requiring external stimulation [1]. Yet MI-EEG decoding remains difficult. The signals are weak, non-stationary, noisy, and highly variable across individuals. These properties make it challenging to distinguish genuine task-relevant structure from subject-specific nuisance structure.

Over the last decade, deep learning has become increasingly prominent in MI-EEG analysis because it can learn highly nonlinear features directly from data [2–4]. Parallel and transformed-

representation models have further shown that spatio-temporal coding, connectivity-aware decoding, and image-like representations can all improve raw benchmark performance when the evaluation protocol is permissive enough [5–7]. However, the field has also inherited a methodological problem that is now too important to ignore: model performance depends not only on the representation or architecture, but also on the unit of data partitioning and the transparency of cohort selection. Recent work on Electroencephalography deep learning indicates that sample-based cross-validation may yield optimistic estimates because subject-specific characteristics can compromise fold independence, while even subject-based validation may remain vulnerable when validation and test partitions are not strictly separated [8]. In short, validation strategy is not a minor implementation detail; it changes the scientific meaning of the reported accuracy.

This problem is especially visible in the MI-EEG literature on the PhysioNet EEG Motor Movement/Imagery dataset [9]. A substantial subset of papers on this dataset has relied on segment-level, image-level, or time-resolved validation [6,10–12], while many others have used fewer than all available volunteers because of annotation problems, missing trials, computational tractability, quality filtering, or deliberate reduced-cohort benchmarking during model development [4,13–16]. None of these practices automatically implies scientific misconduct. Some are understandable, and some are useful at the representation-design stage. But if protocol strength is not distinguished from raw predictive performance, then direct comparison across papers becomes unreliable. Strong reported accuracy may reflect a strong model, an easier split, a smaller and more homogeneous cohort, or some combination of all three.

We recently proposed a GAF–PLV parallel CNN as a proof-of-concept segment-wise benchmarking framework for MI-EEG [7]. In that setting, the representation family achieved 99.73% binary accuracy on a 10-subject subset and 99.18% on a 30-subject subset. Those results established the promise of the spatio-temporal encoding for capturing nonlinear and non-stationary MI-EEG structure. The present study extends that line of work by testing the same framework under a subject-wise inference target, asking how much of that benchmark performance remains when the test subject is wholly unseen and every retained fold is disclosed.

For this reason, the current study adopts leave-one-subject-out cross-validation (LOSOCV) for the binary left-versus-right MI task. The analysis is performed on the 105 held-out subject folds for which subject-wise rerun outputs were available in this study. These folds correspond to subject IDs 1–105. Because the available outputs do not include fold results for subjects 106–109, the present subject-wise study is reported on the retained executable cohort rather than on a hypothetical full 109-subject rerun. This transparency matters: the paper’s scientific claim is tied to the exact folds that are actually available, not to unverified assumptions about missing outputs. The goal is therefore twofold: to provide a stronger estimate of cross-subject performance for the GAF–PLV framework, and to show directly how evaluation regime changes the scientific meaning of performance in MI-EEG research. In methodological terms, the paper aims to extend a strong proof-of-concept benchmark into a more rigorous subject-wise reference against which later claims of subject-independent MI-EEG performance can be judged.

The main contributions of this paper are as follows.

1. We provide a fold-complete subject-wise LOSOCV study of our recently proposed GAF–PLV representation family on the retained executable cohort.
2. We report a transparent retained-cohort analysis on the 105 held-out subject folds for which subject-wise rerun outputs were available in this study and pair it with bootstrap and inferential statistics so that cross-subject uncertainty is visible rather than implicit.
3. We provide a direct protocol-sensitive comparison with our earlier pooled-window proof-of-concept benchmark, thereby showing how the same representation family changes meaning when the validation unit changes from windows to unseen subjects.
4. We quantify that protocol-sensitive shift using both the cohort mean and held-out subject anchors rather than treating the drop as a single fixed number.

- 5. We convert a strong proof-of-concept benchmark result into a more rigorous field reference by making the retained cohort, fold-level variability, uncertainty, and scope boundaries explicit.
- 6. We distil the empirical findings into a practical reporting standard for future subject-independent MI-EEG studies and add exploratory classical-baseline context while making cohort mismatch explicit.

This paper therefore focuses on the scientific impact of protocol change: it shows how a strong proof-of-concept MI-EEG framework behaves when the inference target shifts from benchmark-oriented segment-level discrimination to subject-independent evaluation. Its importance lies in providing a more trustworthy reference for future method development, evaluation, and translational interpretation in MI-EEG.

2. Related Work

MI-EEG classification has generally developed along two broad directions: traditional machine learning pipelines based on handcrafted features, and deep learning pipelines that either learn directly from raw signals or operate on transformed representations [1,2]. Traditional approaches typically combine preprocessing, handcrafted feature extraction, and classification using methods such as LDA, SVM, KNN, or random forests, whereas deep learning has shifted the field toward convolutional, recurrent, and graph-based architectures that aim to learn discriminative features more directly from the data [3–6]. Yet the field’s methodological maturity has lagged behind its architectural creativity.

A major source of that mismatch is data partitioning. As Del Pup *et al.* show, EEG deep learning results can vary substantially depending on whether cross-validation is performed at the sample level, the subject level, or within a nested subject-based design. Sample-based approaches can place segments from the same recording across model-development and evaluation folds, which may inflate apparent accuracy because EEG contains strong subject-specific structure. Even standard subject-based approaches can remain optimistic if the same held-out data are used both to monitor training and to estimate final performance. Their large-scale comparative study therefore argues for subject-based strategies, and more specifically for nested subject-based strategies, whenever unseen-subject generalisation is the target of inference [8]. This point is highly relevant to MI-EEG, where within-subject benchmarking and cross-subject deployment are often conflated.

A separate but related problem is cohort definition. On the PhysioNet motor imagery dataset, many studies use fewer than the full 109 volunteers. In some cases this is due to missing trials, corrupted annotations, or documented data anomalies. In other cases it reflects deliberate reduced-cohort benchmarking, often motivated by computational cost or early-stage model development. Again, this practice is not automatically illegitimate. Indeed, reduced-cohort experimentation has been common on this dataset. The problem arises when cohort selection is not reported transparently, when exclusions are outcome-driven, or when reduced-cohort benchmarking is rhetorically upgraded into a claim of subject-independent superiority.

These two issues – validation protocol and cohort transparency – should therefore be treated explicitly in the literature review rather than hidden in implementation details. Table 1 summarises representative PhysioNet MI studies whose validation profiles should be interpreted with care when the target claim is unseen-subject inference. The point of the table is not to invalidate these papers. It is to clarify how directly each reported result supports cross-subject interpretation and practical subject-independent use.

Table 1. Representative PhysioNet MI studies with validation profiles relevant to cross-subject interpretation. The final column summarises how directly each result supports unseen-subject inference.

Paper	Year	Subjects	Protocol	Main interpretive note	Inference scope
Enhanced classification of MI EEG signals using spatio-temporal representations (our recent proof-of-concept study) [7]	2025	10 and 30	Windows generated before split; 9:1 sample-level split; five-fold sample-level CV	Strong benchmark result under segment-wise evaluation; not designed as a strict unseen-subject study	Benchmark only
Lomelin-Ibarra <i>et al.</i> [10]	2022	105	80% of generated samples for training, 10% for validation, 10% for test over image-like representations	Image-level pooled-sample split	Limited
Roots <i>et al.</i> [11]	2020	103	70% of pooled samples for training, 10% for validation, 20% for testing	Pooled-sample split despite cross-subject framing	Limited
Chowdhury <i>et al.</i> [12]	2023	103	70% random data for training, 10% for validation, 20% for testing	Random pooled splitting rather than strict unseen-subject evaluation	Limited
Huang <i>et al.</i> [17]	2023	single-subject experiments	Recurrence-plot image classification on one subject at a time	Useful for within-subject analysis, but not for broad cross-subject claims	Limited
Ghimire and Sekeroglu [18]	2022	109	Subject-block style validation using subsets of subjects for validation under a global setting	More structured than pooled random splitting, but not a clean LOSO or nested unseen-subject design	Partial
Huang <i>et al.</i> [19]	2022	109	Global classifier evaluated by grouped subject partitions; individual variability also examined	Cleaner than sample-level splits, but accessible protocol remains insufficiently explicit for a full LOSO interpretation	Partial

Protocol heterogeneity is only one side of the problem. The other is selective or reduced cohort usage on the same dataset. Table 2 summarises representative studies that used fewer than all 109 PhysioNet volunteers. The important point here is subtle. The literature shows that reduced-cohort experimentation has been common, but this does not make it methodologically optimal. What matters is why fewer subjects were used, whether the rule was explicit, and whether the retained cohort was then analysed comprehensively or selectively.

Table 2. Representative same-dataset PhysioNet MI studies using fewer than all 109 volunteers. The table is illustrative rather than exhaustive and is included to document heterogeneous subject-count practices in the EEGMMIDB literature.

Study	Year	Subjects	Why fewer than 109 were used	Interpretation
Dose <i>et al.</i> [4]	2018	105	Subset of 105 subjects retained due to missing trials in the public dataset	Criterion-based exclusion
Lun <i>et al.</i> [13]	2020	10, 20, 60, 100	Deliberate subgroup benchmarking on the PhysioNet dataset	Reduced-cohort benchmarking
Roots <i>et al.</i> [11]	2020	103	Six subjects omitted because of incorrectly annotated data	Criterion-based exclusion
Wang <i>et al.</i> [20]	2020	105	Four subjects discarded because of variability in number of trials	Criterion-based exclusion
Li <i>et al.</i> [14]	2021	48 of 104	Quality-based outlier filtering before experimentation	Quality-filtered subset
Majoros and Oniga [15]	2022	10 and 20	Deliberate reduced-cohort benchmarking on the PhysioNet four-class dataset	Reduced-cohort benchmarking
Hwaidi and Chen [21]	2022	10	Deliberate 10-subject PhysioNet experiment	Reduced-cohort benchmarking
Fan <i>et al.</i> [22]	2023	20	Selected 20 PhysioNet subjects for evaluation/validation	Reduced-cohort benchmarking
Aung <i>et al.</i> [16]	2025	20	Empirical study conducted on 20 PhysioNet subjects	Reduced-cohort benchmarking

Within this broader landscape, the present paper takes a narrower and more specific position. We do not argue that early-stage reduced-cohort benchmarking is without value; indeed, it helped establish the recently proposed GAF–PLV representation. Our point is that when the target claim is cross-subject transfer, protocol choice and retained-cohort disclosure must be explicit. Our own study therefore uses a simple rule: analyse the retained executable cohort transparently, report all available LOSOCV folds, and interpret performance in light of the stronger inference target. This is the methodological stance on which the rest of the paper is built. Segment-wise and pooled-sample studies can still be valuable for rapid representation screening, ablation testing, and early benchmark development. They are not, however, practical proxies for plug-and-play subject-independent MI-BCI, because they do not test whether a trained model transfers to a wholly unseen individual. What the literature still lacks, and what the present paper seeks to supply, is a transparent subject-wise reference object whose scientific meaning is difficult to misread and whose methodological lessons can be reused by later studies.

3. Methodology

3.1. Study Design and Rationale

The present study is a subject-independent evaluation of our recently proposed GAF–PLV spatio-temporal MI-EEG framework. The central representational idea is retained: a temporal branch based on GAF-derived channel correlation and a spatial branch based on PLV-derived functional connectivity are decoded jointly by a dual-input CNN. What changes is not the descriptor family but the inference target. Rather than treating pooled windows as the unit of validation, the present study uses a subject-wise LOSOCV design so that every reported test fold corresponds to an unseen individual. The resulting evidence therefore speaks directly to cross-subject transfer and to the size of the generalization gap between segment-level benchmarking and unseen-subject testing. More specifically, the study is designed to answer three tightly related questions: how large the protocol-sensitive performance drop is when evaluation moves from pooled windows to unseen subjects, whether the representation retains

any above-chance transferable signal under that stricter test, and how much subject-wise heterogeneity is hidden by a single cohort mean.

3.2. Dataset and Problem Formulation

The experiments are based on the PhysioNet motor imagery EEG dataset acquired with 64 electrodes at a sampling rate of 160 Hz [9]. The underlying temporal and spatial feature-construction stages follow our recently proposed GAF-PLV pipeline, whereas the present paper changes the evaluation regime from proof-of-concept pooled-window benchmarking to subject-wise LOSOCV [7]. Figure 1 now provides the consolidated end-to-end workflow used in the present paper. Reading from left to right, the figure first defines the retained cohort and the binary task, then shows 8–30 Hz filtering, CAR, per-trial z-scoring, and segmentation of each 4 s trial into ten non-overlapping 0.4 s windows. The upper branch constructs one GAF matrix per channel and integrates the 64 channel-wise GAF descriptors through Spearman correlation to form the temporal image, whereas the lower branch computes the PLV-based spatial image from the same window. Both descriptors are then resized to $256 \times 256 \times 3$ before being processed by matched CNN branches. The right-hand side of Figure 1 also clarifies that the shown window-level class outputs are illustrative, that the final trial decision is obtained by the mode of the ten window predictions, and that the available evaluation uses a subject-wise LOSOCV design with one held-out test subject and one separate validation subject per outer fold. Each MI trial lasts 4 s and therefore contains 640 temporal samples per channel. In the currently available study outputs, the subject-wise analysis is binary and uses 105 valid subjects. Let the EEG trial from subject s , trial r , channel c , and time index t be denoted by

$$x_{s,r,c}(t), \quad s \in \mathcal{S}, r \in \mathcal{R}, c \in \{1, \dots, 64\}, t \in \{1, \dots, 640\}. \quad (1)$$

The binary label is

$$y_{s,r} \in \{0, 1\}, \quad (2)$$

where class 0 denotes left-hand MI and class 1 denotes right-hand MI.

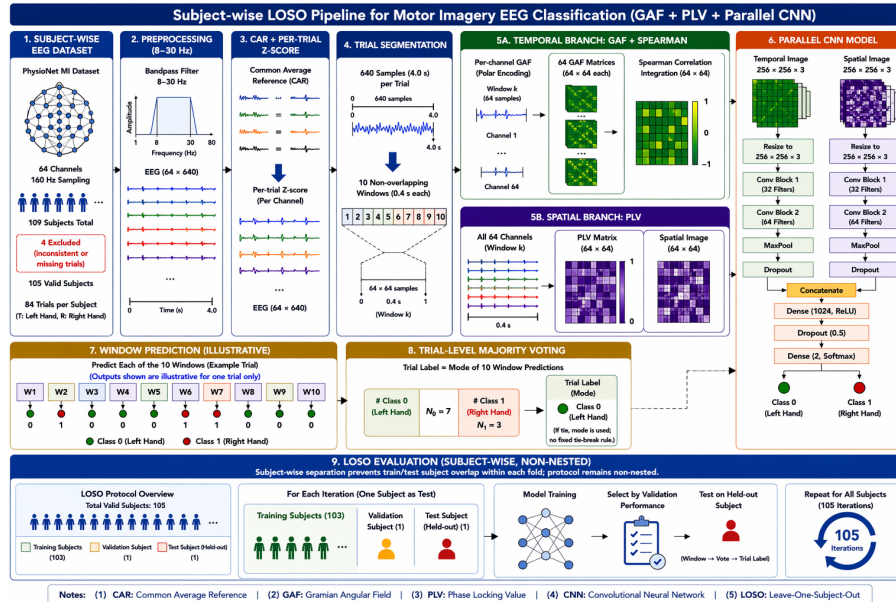


Figure 1. Integrated workflow of the binary subject-wise LOSOCV study. The figure consolidates the retained-cohort definition (109 total subjects, 105 valid after four data-integrity exclusions), 8–30 Hz preprocessing, CAR and per-trial z-scoring, ten non-overlapping 0.4 s windows per 4 s trial, temporal GAF–Spearman encoding, spatial PLV encoding, resizing to $256 \times 256 \times 3$ image inputs, dual-branch CNN decoding, illustrative window-level predictions, trial-level majority voting, and the subject-wise LOSOCV protocol with separate training, validation, and held-out test subjects in each outer fold.

3.3. Subject Inclusion, Exclusion, and Reporting Transparency

The official PhysioNet EEG Motor Movement/Imagery dataset contains recordings from 109 volunteers [9]. However, the available binary LOSOCV rerun outputs used in this study contain subject-wise fold results only for subject IDs 1–105. In other words, the retained executable cohort is known exactly, whereas fold outputs for subjects 106–109 are absent from the available study records. We therefore report the study on the basis of what is verifiably available: 105 held-out subject folds, all of which are included in the analysis. This is stricter than inferring undocumented exclusions from memory or from narrative convenience. It also prevents a common weakness in EEG reporting, namely, quietly changing the effective cohort without showing the exact fold-level evidence. Accordingly, the present paper does not claim more than the files support. It claims a transparent LOSOCV study on the 105 retained subject folds for which subject-wise outputs are available, and every one of those folds contributes to the reported summary metrics and figures.

3.4. Preprocessing and Temporal Partitioning

Consistent with the earlier GAF–PLV pipeline, each trial is filtered in the 8–30 Hz motor-rhythm band before common average reference is applied independently at each time point:

$$\tilde{x}_{s,r,c}(t) = x_{s,r,c}(t) - \frac{1}{C} \sum_{k=1}^C x_{s,r,k}(t), \quad C = 64. \quad (3)$$

Each channel is then z-scored independently within the trial:

$$z_{s,r,c}(t) = \frac{\tilde{x}_{s,r,c}(t) - \mu_{s,r,c}}{\sigma_{s,r,c} + \varepsilon}, \quad (4)$$

with

$$\mu_{s,r,c} = \frac{1}{T} \sum_{t=1}^T \tilde{x}_{s,r,c}(t), \quad (5)$$

$$\sigma_{s,r,c} = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (\tilde{x}_{s,r,c}(t) - \mu_{s,r,c})^2}, \quad (6)$$

where $T = 640$.

Each 4 s trial is divided into ten non-overlapping windows of length 0.4 s. Since the sampling rate is 160 Hz,

$$L = 0.4 \times 160 = 64, \quad (7)$$

so each trial contains

$$W = \frac{640}{64} = 10 \quad (8)$$

windows. The w -th window is denoted by $\mathbf{Z}_{s,r}^{(w)} \in \mathbb{R}^{L \times C}$. The ten-window partitioning is a fixed preprocessing rule used for every trial before temporal and spatial image construction.

3.5. Temporal Representation Using GAF and Spearman Integration

For a channel signal $\mathbf{u}_c = [u_c(1), u_c(2), \dots, u_c(L)]^\top$, the temporal descriptor is built using the Gramian Angular Field formulation of Wang and Oates, followed by the same channel-integration strategy used in the earlier GAF–PLV pipeline [23]. Min–max normalisation to $[-1, 1]$ is performed as

$$u'_c(i) = \begin{cases} 2 \frac{u_c(i) - \min(\mathbf{u}_c)}{\max(\mathbf{u}_c) - \min(\mathbf{u}_c)} - 1, & \max(\mathbf{u}_c) \neq \min(\mathbf{u}_c), \\ 0, & \max(\mathbf{u}_c) = \min(\mathbf{u}_c). \end{cases} \quad (9)$$

The angular encoding is

$$\phi_c(i) = \arccos(u'_c(i)), \quad i = 1, \dots, L, \quad (10)$$

and the Gramian Angular Field matrix for channel c is

$$\mathbf{G}_c(i, j) = \cos(\phi_c(i) + \phi_c(j)), \quad i, j = 1, \dots, L. \quad (11)$$

Since one GAF matrix is produced per channel, a given window initially yields 64 GAF matrices. These are integrated into a single temporal summary matrix by Spearman correlation between vectorised GAF matrices:

$$\mathbf{T}(c, d) = \rho_s(\text{vec}(\mathbf{G}_c), \text{vec}(\mathbf{G}_d)), \quad c, d = 1, \dots, 64. \quad (12)$$

The resulting temporal image descriptor $\mathbf{T} \in \mathbb{R}^{64 \times 64}$ is resized and stored as a $256 \times 256 \times 3$ image. Figure 2 illustrates this construction for Subject 22: representative single-channel GAF matrices are shown together with the final integrated Spearman matrix that serves as the temporal branch input to the CNN.

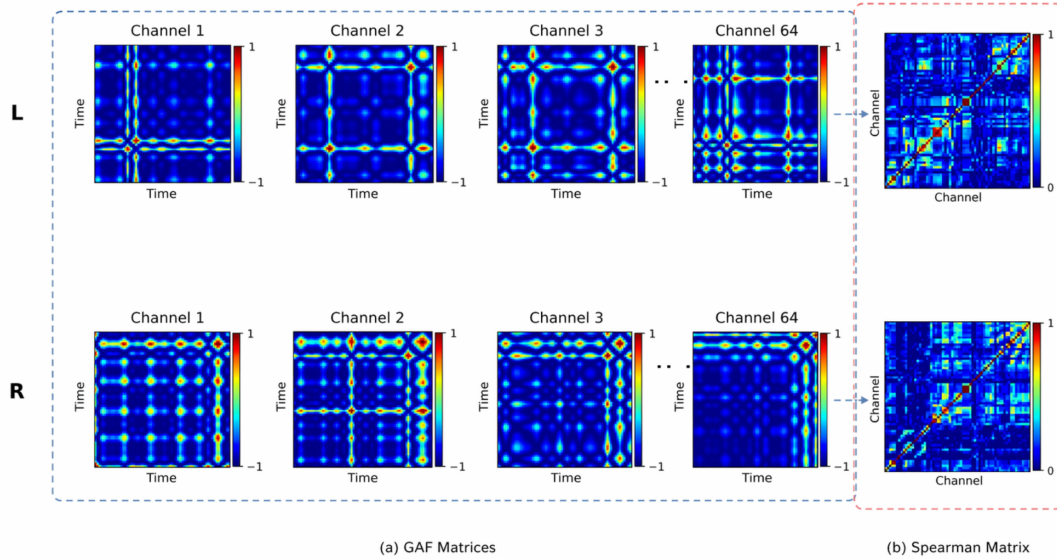


Figure 2. Representative temporal-image construction for Subject 22 in the binary left-versus-right task. For each class, channel-wise GAF matrices are computed for the selected 0.4 s window, and the final integrated Spearman matrix summarises cross-channel relationships between the vectorised GAF descriptors.

3.6. Spatial Representation Using PLV

For the same window, the analytic signal of channel c is computed by the Hilbert transform:

$$a_c(t) = u_c(t) + j\mathcal{H}\{u_c(t)\}, \quad (13)$$

with instantaneous phase

$$\theta_c(t) = \arg(a_c(t)). \quad (14)$$

For channels c and d , the phase difference is

$$\Delta\theta_{c,d}(t) = \theta_c(t) - \theta_d(t). \quad (15)$$

The phase-locking value is then

$$\text{PLV}(c, d) = \left| \frac{1}{L} \sum_{t=1}^L e^{j\Delta\theta_{c,d}(t)} \right|. \quad (16)$$

This yields the spatial connectivity matrix $\mathbf{P} \in \mathbb{R}^{64 \times 64}$, which is also resized and stored as a $256 \times 256 \times 3$ image. Following the earlier GAF-PLV pipeline, the PLV matrix is binarised using the upper quartile of off-diagonal weights so that the functional graph retains only stronger synchronisation edges. Figures 3, 4, and 5 show the corresponding PLV matrices, their binarised thresholded forms, and the resulting functional brain graphs for representative left- and right-imagery windows from Subject 22.

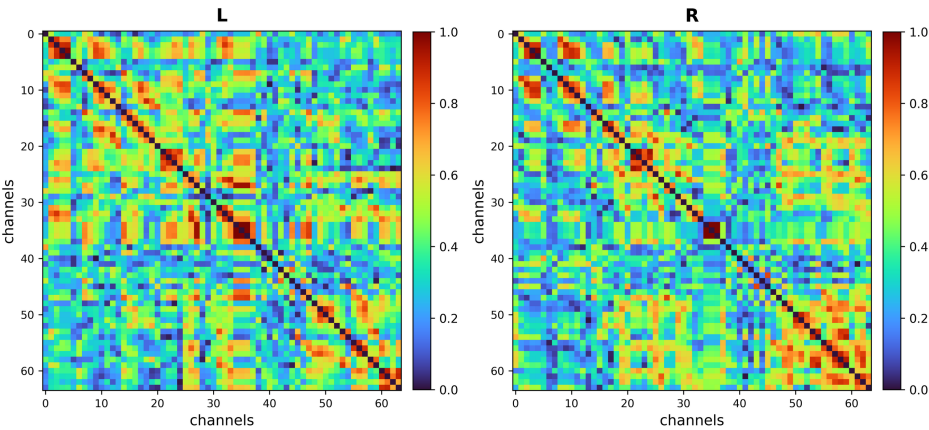


Figure 3. PLV matrices for representative left- and right-imagery windows from Subject 22. These matrices quantify phase synchronisation between channel pairs before thresholding.

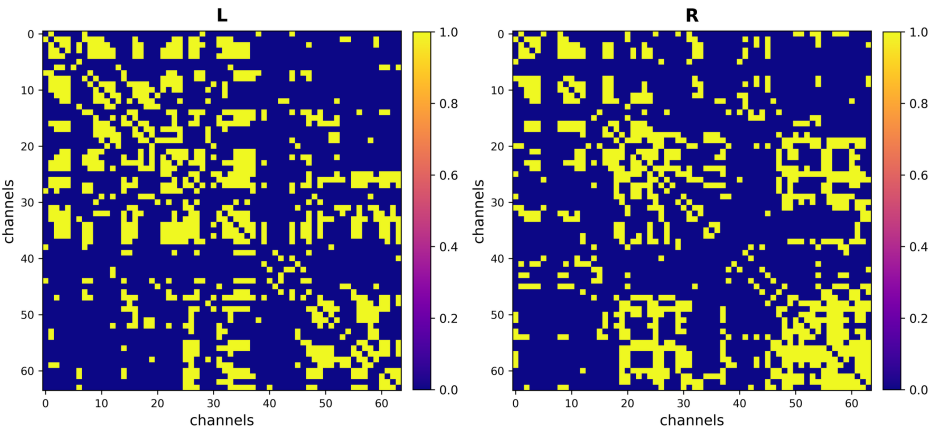


Figure 4. Binarised PLV matrices for Subject 22 after applying the upper-quartile threshold to the off-diagonal PLV weights. This step suppresses weak connections and retains stronger candidate edges for graph construction.

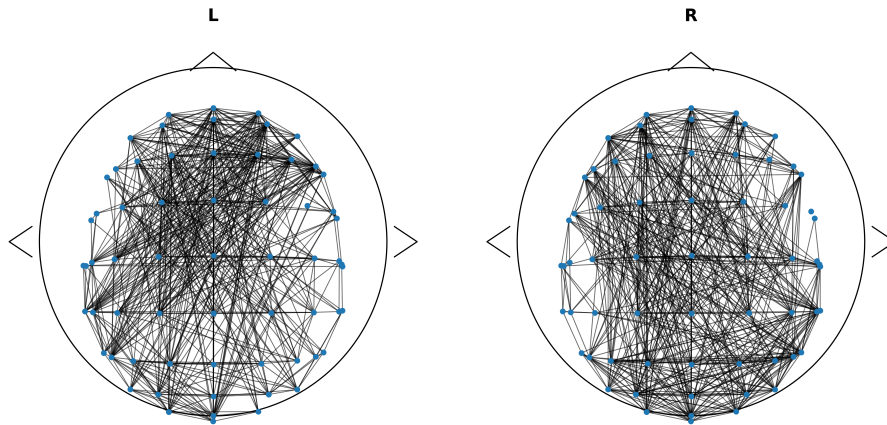


Figure 5. Functional brain networks obtained from the thresholded PLV matrices for Subject 22. Although these plots are illustrative rather than inferential, they show how the spatial branch converts synchronisation structure into an interpretable graph representation before image-based CNN decoding.

3.7. Dual-Input Parallel CNN

The temporal and spatial descriptors are decoded by a dual-input parallel CNN that follows our recently proposed GAF-PLV model family, interpreted here under a subject-wise LOSOCV regime rather than a pooled-window regime. As summarised in Figure 1, the descriptors used by the executable pipeline are resized image inputs; however, the layer-wise schematic in Figure 6 is drawn at the native 64×64 descriptor scale in order to show the branch architecture, kernel orientation, channel expansion, and fusion stages more explicitly.

The temporal branch uses asymmetric early convolutions with 1×2 kernels, whereas the spatial branch uses complementary 2×1 kernels. This asymmetry is intended to bias the first stage of feature extraction toward directional local structure that differs between the GAF-based temporal representation and the PLV-based spatial representation. In both branches, three initial convolutions expand the feature depth from 32 to 64 channels before the first max-pooling stage. Two deeper 3×3 convolutions then increase the representation to 256 channels, followed by a second pooling stage and a final 3×3 convolution with 256 channels before the third pooling stage. According to the schematic in Figure 6, each branch is then flattened to a 9216-dimensional vector and projected through a branch-specific fully connected layer with 1024 units.

The two 1024-dimensional branch embeddings are concatenated into a 2048-dimensional fusion vector,

$$\mathbf{h} = [\mathbf{h}^{(T)}; \mathbf{h}^{(S)}], \quad (17)$$

which is then passed to a shared fully connected layer with 512 units before final softmax classification:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}\mathbf{h} + \mathbf{b}). \quad (18)$$

Training uses sparse categorical cross-entropy,

$$\mathcal{L} = - \sum_{k=0}^1 \mathbb{K}(y = k) \log \hat{y}_k, \quad (19)$$

with the Adam optimiser, maximum 100 epochs, batch size 64, and learning rate 10^{-5} . Figure 6 should therefore be read as a compact layer-by-layer architectural explanation of the branch design, while Figure 1 remains the authoritative end-to-end summary of how descriptor construction, resizing, branch decoding, and subject-wise evaluation interact in the present study.

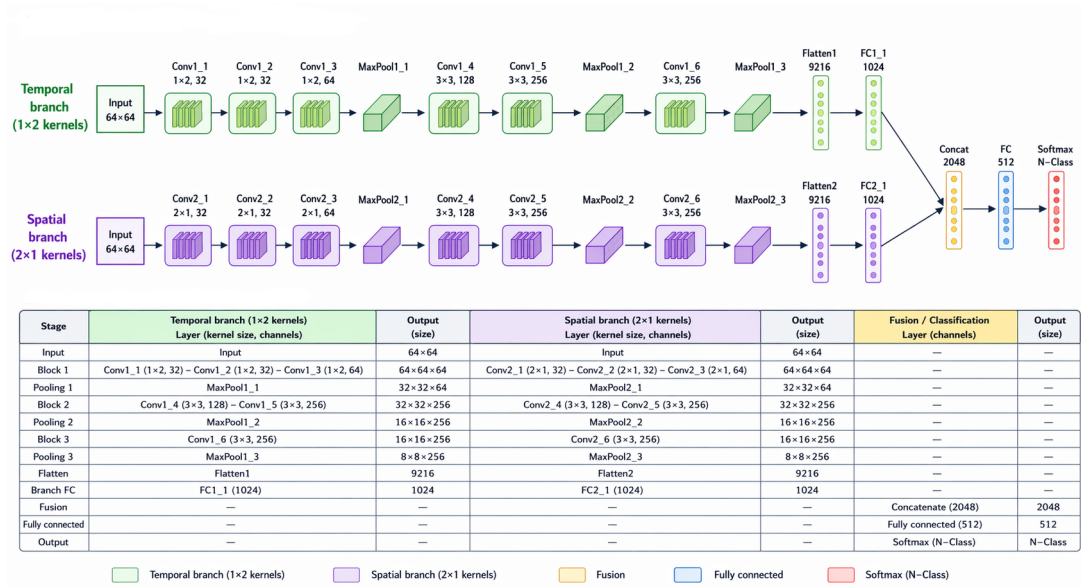


Figure 6. Detailed dual-input parallel CNN architecture used for binary LOSOCV decoding. The temporal branch applies early 1×2 convolutions and the spatial branch applies early 2×1 convolutions before deeper 3×3 convolutional blocks, branch-specific flattening and 1024-unit fully connected layers, 2048-dimensional late fusion by concatenation, a 512-unit shared fully connected layer, and final softmax classification. The schematic is adapted to show the branch structure explicitly at the native 64×64 descriptor scale.

3.8. Subject-Wise LOSOCV

Let $\mathcal{S} = \{1, 2, \dots, N\}$ denote the set of valid subjects with $N = 105$. For each outer fold, one subject s^* is held out for testing:

$$\mathcal{S}_{\text{test}} = \{s^*\}. \quad (20)$$

A different subject from the remaining pool is chosen for validation, and the rest are used for training:

$$\mathcal{S}_{\text{train}} = \mathcal{S} \setminus (\mathcal{S}_{\text{test}} \cup \mathcal{S}_{\text{val}}). \quad (21)$$

Thus, one subject appears only in train or validation or test, never in more than one split in the same fold.

Window-level predictions are aggregated back to the trial level by majority vote, exactly as visualised in Figure 7. The schematic shows the split into training, validation, and test subjects together with ten explicit window-level predictions and the resulting trial-level majority-vote rule:

$$\hat{y}_{s,r} = \text{mode}\{\hat{y}_{s,r,1}, \hat{y}_{s,r,2}, \dots, \hat{y}_{s,r,10}\}. \quad (22)$$

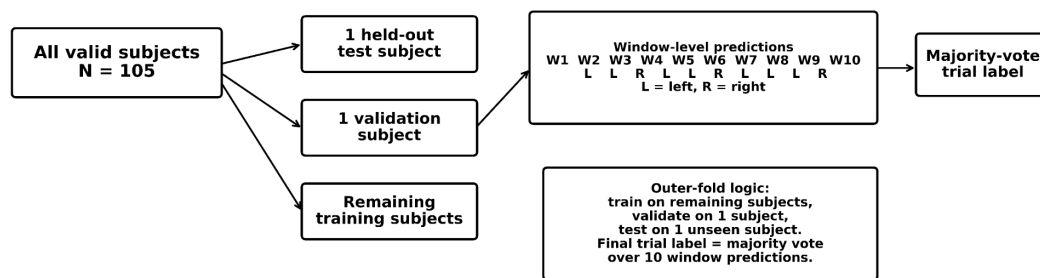


Figure 7. Subject-wise LOSOCV protocol and trial-level majority voting used in the present study. In each outer fold, one subject is reserved for testing, one different subject is used for validation, and the remaining subjects are used for training. Ten window-level predictions for a representative trial are shown explicitly inside the prediction block, where L and R denote left- and right-hand motor imagery predictions, respectively, and the final trial label is obtained by majority vote over the ten outputs.

3.9. Performance Metrics

At the trial level, the following metrics are reported:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (23)$$

$$\text{Balanced Accuracy} = \frac{1}{2} \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right), \quad (24)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \quad (25)$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (26)$$

Cohen's kappa is computed as

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad (27)$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

4. Results and Discussion

4.1. Cohort-Level LOSOCV Performance

Across the 105 retained held-out subject folds, the binary subject-wise study produced a mean trial-level accuracy of $58.07\% \pm 8.27\%$, Macro-F1 of $53.48\% \pm 11.19\%$, macro precision of $60.97\% \pm 14.45\%$, and Cohen's kappa of 0.1615 ± 0.1654 . The median accuracy was 57.14% , the interquartile range was 52.38% to 64.29% , the minimum was 38.10% , and the maximum was 78.57% . These values show that the central empirical result is not a single headline score but the breadth of fold-to-fold behaviour once the held-out subject is fully excluded from model fitting.

These cohort-level values matter because they quantify performance after subject recurrence has been removed from the evaluation logic. In a protocol-sensitive literature, a lower but better-specified estimate is more informative for cross-subject inference than a near-ceiling result obtained under pooled partitions [8,10–12]. For real subject-independent use, the latter should be read as benchmark-oriented results rather than as deployment-ready performance estimates. Figures 8, 9 and 10 should therefore be read together: the first condenses the non-redundant cohort-level metrics, the second shows the full held-out subject accuracy profile, and the third visualises the subject-wise behaviour of Macro-F1 and kappa.

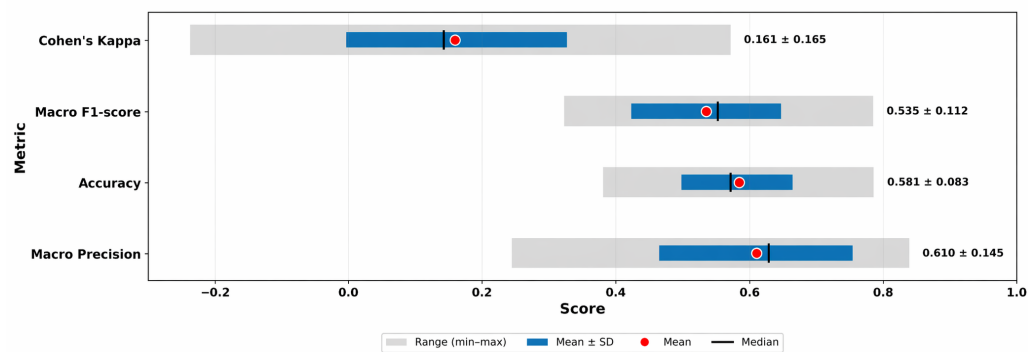


Figure 8. Overall summary of the non-redundant LOSOCV metrics. For each metric, the light bar shows the observed min–max range, the dark bar shows mean ± standard deviation, the dot shows the mean, and the central tick shows the median. Balanced accuracy and macro recall are omitted because they are numerically identical to accuracy in the present balanced binary evaluation.

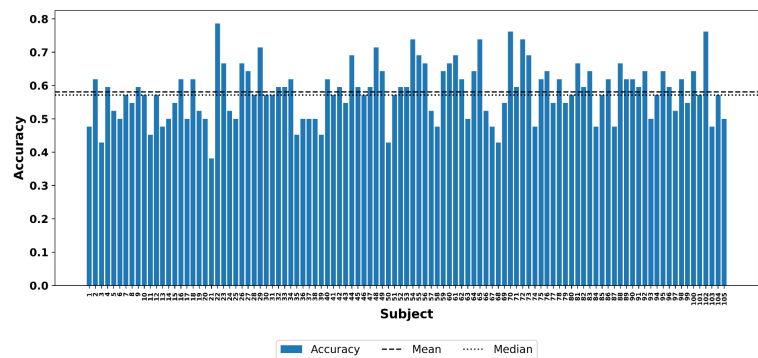


Figure 9. Subject-wise LOSOCV accuracy across all 105 retained folds in subject order. The dashed and dotted reference lines indicate the cohort mean and median, respectively, so that fold-level variability can be read directly against the central tendency.

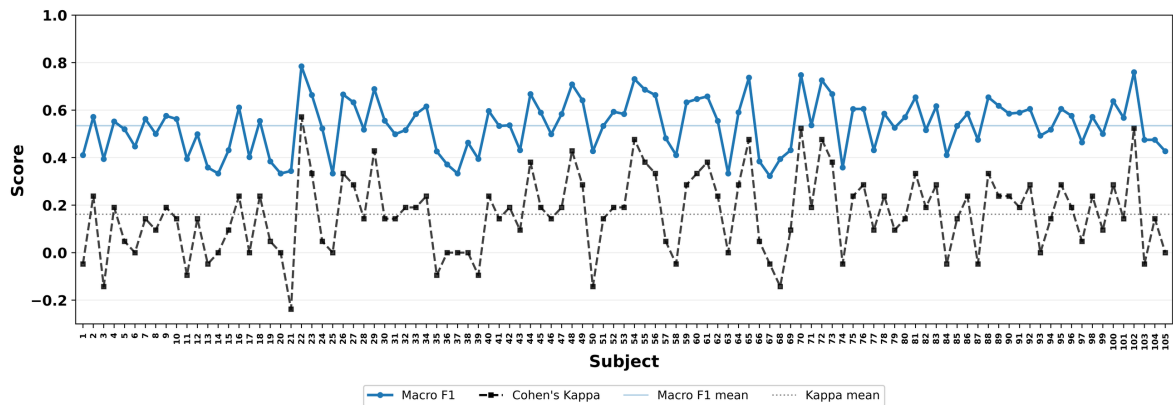


Figure 10. Subject-wise Macro-F1 and Cohen’s kappa across all 105 LOSOCV folds. Macro-F1 remains consistently below accuracy, while kappa is markedly lower and occasionally negative, highlighting limited chance-corrected agreement for a subset of held-out subjects.

4.2. Statistical Analysis of Subject-Wise Performance

To make the subject-wise LOSOCV interpretation less impressionistic, statistical analysis was performed directly on the 105 held-out subject folds. For each metric, we report mean, standard deviation, median, quartiles, minimum, maximum, and a 95% bootstrap confidence interval (CI) of the mean obtained from 20,000 bootstrap resamples. Normality of the subject-wise distributions was examined with the Shapiro–Wilk test. Accuracy and Macro-F1 were then tested against the balanced

binary baseline of 0.5, while Cohen’s kappa was tested against 0. One-sample *t*-tests, Wilcoxon signed-rank tests, and Cohen’s *d* were used so that both parametric and nonparametric evidence were available.

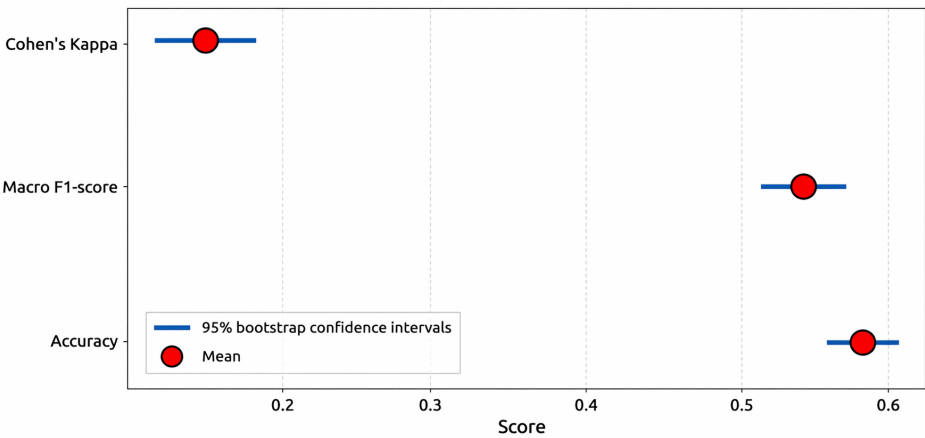


Figure 11. Mean performance with 95% bootstrap confidence intervals for the three non-redundant cohort-level metrics. The confidence intervals do not cross the relevant baselines for accuracy, Macro-F1, or kappa, which is consistent with the inferential tests in Table 3.

Table 3. Subject-wise descriptive and inferential statistics across the 105 LOSOCV folds. CI denotes the 95% bootstrap confidence interval of the mean.

Metric	Mean ± SD	Median	Q1–Q3	Min–Max	95% CI	Baseline	<i>t</i> (104)	Wilcoxon <i>p</i>	Cohen’s <i>d</i>
Accuracy	0.5807 ± 0.0827	0.5714	0.5238–0.6429	0.3810–0.7857	0.5649–0.5964	0.5	10.00 (<i>p</i> < 10 ^{−16})	2.50 × 10 ^{−13}	0.976
Macro-F1	0.5348 ± 0.1119	0.5524	0.4312–0.6050	0.3226–0.7846	0.5139–0.5562	0.5	3.19 (<i>p</i> = 0.0019)	2.27 × 10 ^{−3}	0.311
Cohen’s kappa	0.1615 ± 0.1654	0.1429	0.0476–0.2857	−0.2381–0.5714	0.1306–0.1927	0	10.00 (<i>p</i> < 10 ^{−16})	5.84 × 10 ^{−14}	0.976

The statistical results sharpen the descriptive interpretation. Accuracy was significantly above the balanced binary baseline of 0.5 under both parametric and nonparametric testing, with a large effect size. Macro-F1 was also significantly above 0.5, but with a much smaller effect size, which indicates that class-balanced performance is materially weaker than raw correctness. Subject-wise accuracy was significantly higher than subject-wise Macro-F1 (mean difference 0.0459; paired *t*(104)=10.23, *p* < 10^{−16}; Wilcoxon *p* = 2.91 × 10^{−19}; Cohen’s *d*_z = 0.999), confirming that accuracy alone paints an overly favourable picture of stability. Cohen’s kappa was significantly above zero, yet its magnitude remained low and negative values were observed for a subset of subjects, which is consistent with weak chance-corrected agreement in the lower-performing tail. In the present balanced binary setting, kappa was also a deterministic affine transform of accuracy ($\kappa = 2 \times \text{Accuracy} - 1$ to machine precision), so it should be interpreted as a chance-corrected restatement of the same subject-level pattern rather than as wholly independent evidence.

4.3. Subject-to-Subject Variability

To make the subject-wise behaviour explicit, this paper reports the full held-out subject profile rather than only a cohort mean. The held-out subject profile reveals a wide spread between the lowest-performing fold (Subject 21: 38.10% accuracy, 34.38% Macro-F1, −0.238 kappa), a representative central fold (Subject 7: 57.14% accuracy, 56.25% Macro-F1, 0.143 kappa), and the strongest retained fold (Subject 22: 78.57% accuracy, 78.46% Macro-F1, 0.571 kappa). Figure 9 shows the full unsorted subject-wise pattern, while Figure 10 shows the corresponding Macro-F1 and kappa traces. These

three points are reported as anchors within a heterogeneous cohort so that the breadth of cross-subject behaviour is visible alongside the cohort summary.

The LOSOCV results therefore show substantial heterogeneity across individuals. This variability is visible directly in Figures 8, 9, and 10. Some held-out subjects are classified reasonably well, whereas others remain close to chance. Quantitatively, 90 of 105 subjects (85.7%) are at or above 50% accuracy, 39 subjects (37.1%) reach at least 60%, only 8 subjects (7.6%) reach at least 70%, and just 3 subjects (2.9%) exceed 75%. At the lower end, 15 subjects (14.3%) fall below 50% accuracy, and one subject falls below 40%. This heterogeneity is not a nuisance detail; it is one of the main scientific results of the paper.

The same conclusion appears in the companion metrics. Macro-F1 ranges from 32.26% to 78.46%, and kappa ranges from -0.238 to 0.571 . Thus, the model’s errors are not merely random noise around a single mean. Rather, there is a structured subject-wise dispersion that reflects genuine instability of cross-subject transfer. This observation is consistent with the broader EEG partitioning literature. Subject-specific properties of EEG can be learned and exploited by deep networks, which is precisely why subject-based validation is necessary when the target claim is cross-subject generalisation rather than within-subject adaptation [8]. The present fold-by-fold spread therefore supports a cautious interpretation: the proposed representation retains some cross-subject signal, but it is not yet sufficiently invariant to individual EEG idiosyncrasies.

4.4. Class-Wise Performance

Class-wise performance is shown in Figure 12. The left class achieved precision 0.5711, recall 0.6481, and F1-score 0.6072, whereas the right class achieved precision 0.5933, recall 0.5134, and F1-score 0.5504. The main asymmetry therefore lies in recall: left-hand imagery is recovered more reliably than right-hand imagery, whereas right-hand precision is only slightly higher.

This imbalance is not catastrophic, but it reinforces the wider subject-wise instability visible in Figures 9 and 10. A deployment-oriented subject-independent MI-EEG model would ideally be both stronger overall and more symmetric across the two motor states. The explicit class-wise profile therefore helps localise where further gains in cross-subject robustness are most likely to arise, particularly in right-hand recall.

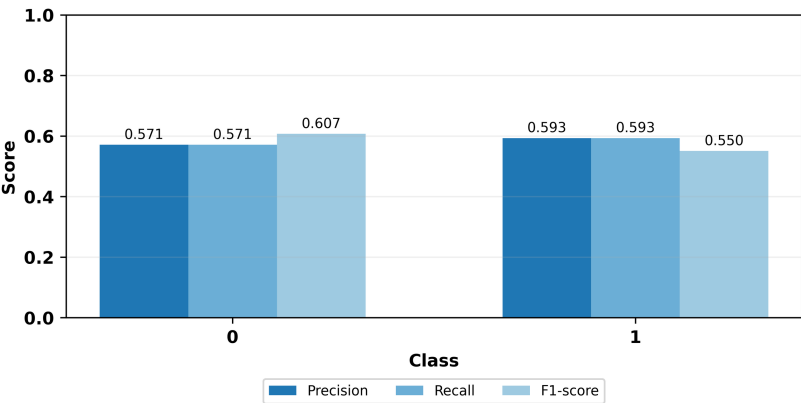


Figure 12. Class-wise binary precision, recall, and F1-score under LOSOCV. The main asymmetry lies in recall, where left-hand imagery is recovered more reliably than right-hand imagery.

4.5. Protocol Sensitivity Relative to the Earlier Pooled-Window Proof-of-Concept Study

The present paper keeps the underlying GAF-PLV feature construction and parallel-CNN decoding logic, but changes the inference target. In the earlier proof-of-concept benchmark, 0.4 s windows were generated before data partitioning, reduced cohorts of 10 and 30 subjects were used, and a 9:1 sample-level split with five-fold sample-level cross-validation was reported for the binary task. The present study instead evaluates 105 retained subjects under LOSOCV with one fully held-out subject per outer fold. Table 4 summarises that contrast.

These rows are not intended as a like-for-like leaderboard because cohort size, task scope, and split logic differ. They are, however, highly informative about protocol sensitivity. When the same GAF–PLV representation family is moved from pooled-window benchmarking to unseen-subject evaluation, the apparent performance landscape changes materially. Using the 99.73% proof-of-concept benchmark as the common anchor, the mean generalization gap to the retained-cohort LOSOCV result is 41.66 points. Crucially, that shift is not one fixed number across the cohort: it is 21.16 points for the best held-out subject (78.57%), 42.59 points for the median held-out subject (57.14%), and 61.63 points for the lowest-performing held-out subject (38.10%). Reporting those anchors alongside the cohort mean makes the protocol-sensitive behaviour of the framework much more interpretable.

Table 4. Protocol sensitivity of the GAF–PLV representation family on PhysioNet. The first two rows come from the earlier pooled-window proof-of-concept study; the present row reports the retained-cohort LOSOCV analysis. The table is not a like-for-like leaderboard because cohort size, task scope, and split logic differ. It is included to show how much the scientific interpretation changes when the evaluation unit changes.

Setting	Subjects	Task	Partition unit	Validation logic	Reported result	What the result supports
Earlier pooled-window proof-of-concept benchmark	10	2-class	0.4 s windows generated before split	9:1 sample-level split; five-fold sample-level CV	99.73% accuracy	Strong pooled-window discrimination on a reduced cohort; established proof-of-concept representational promise rather than an unseen-subject claim.
Earlier pooled-window proof-of-concept benchmark	30	2-class	0.4 s windows generated before split	9:1 sample-level split; five-fold sample-level CV	99.18% accuracy	Same proof-of-concept message under a larger reduced cohort; still not a strict unseen-subject evaluation.
This work	105	2-class	Whole subject held out at test time; trial-level majority voting	LOSOCV with a separate validation subject in each outer fold	58.07% ± 8.27% mean accuracy	Unseen-subject performance estimate with full retained-fold disclosure and explicit inter-subject variability.

4.6. Why Subject-Wise Evaluation Reveals the Generalization Gap

The drop from the earlier segment-level proof-of-concept benchmark to the present LOSOCV result should not be interpreted as evidence that the representation is useless. Rather, it indicates that the evaluation target has become much harder and much more realistic. Several plausible explanations can account for this gap. First, the network can learn subject-specific EEG signatures that help discrimination when pooled windows from the same subject recur across development and evaluation, but those signatures do not transfer reliably to a wholly unseen individual. Second, although the GAF and PLV descriptors are physiologically motivated, they are not automatically invariant to inter-subject differences in anatomy, rhythm expression, noise profile, and task execution style. Third, non-overlapping windows extracted from the same trial are correlated observations rather than independent experimental units, so permissive window-level splitting can overstate apparent generalizability even when the representation itself is informative.

These explanations are consistent with the wider literature on EEG partitioning and with the fold-by-fold heterogeneity seen in the present paper. The key point is therefore not simply that the percentage drops, but that the scientific meaning of the number changes. Figure 13 makes this visible in two ways: the cohort mean falls by 41.66 points, and the held-out subject gaps span 21.16–61.63 points. Under pooled-window evaluation, the result primarily reflects benchmark-oriented discrimination under a permissive partition. Under subject-wise LOSOCV, the remaining performance is a direct

estimate of how much transferable structure the representation retains once the held-out individual is fully excluded from model fitting.

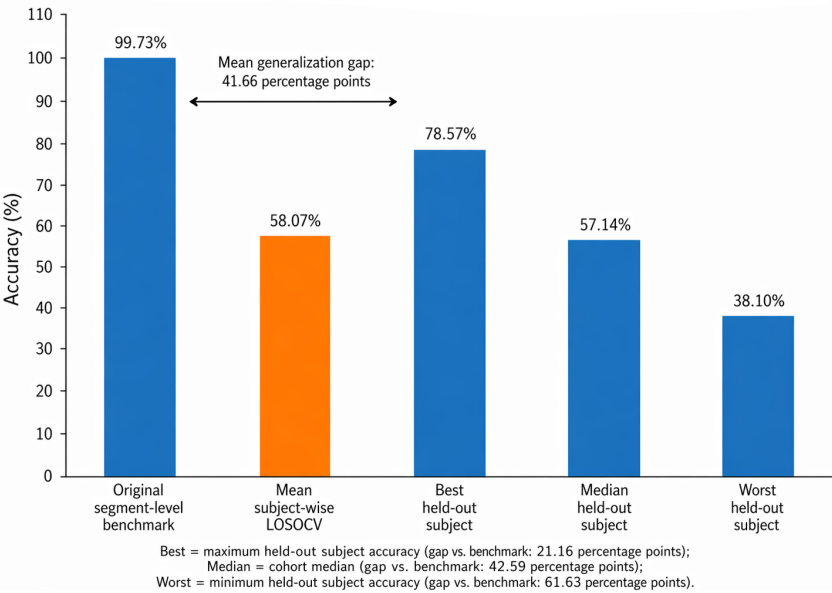


Figure 13. Protocol-sensitive comparison between the earlier segment-level proof-of-concept benchmark and the present subject-wise LOSOCV analysis. The main arrow shows the mean generalization gap of 41.66 points. Relative to the same 99.73% benchmark anchor, the best-, median-, and worst-held-out subject gaps are 21.16, 42.59, and 61.63 points, respectively. The comparison is not like-for-like, but it makes the change in interpretive meaning visually explicit.

4.7. Protocol Trust and the Interpretation of Performance

A key contribution of this paper is methodological rather than architectural. Its originality lies in the evaluation logic and reporting transparency rather than in a new model block. The present LOSOCV results should therefore be read alongside the earlier pooled-window proof-of-concept benchmark and the broader literature showing that validation design can dominate the apparent ranking of MI-EEG models [8,10–12]. The gap between segment-level benchmarking and the present LOSOCV result is not just a change in points; it is a change in the meaning of the evaluation. Under segment-, image-, or time-resolved partitioning, a network may benefit from recurring subject-specific structure across folds. Under LOSOCV, the test subject is absent from model fitting and checkpoint selection. Any performance that remains must therefore come from structure that generalises across individuals rather than from subject recurrence.

For that reason, the present manuscript should be read as a study of generalization rather than as a simple leaderboard exercise. Its main contribution is to show how representational promise changes once the evaluation target becomes the unseen subject. The GAF–PLV representation remains physiologically motivated and interpretable: GAF captures nonlinear temporal correlation and PLV captures synchronisation structure between electrodes [23]. However, the LOSOCV results show that these properties alone do not ensure robust cross-subject transfer. The practical implication is straightforward: a representation that performs extremely well under pooled-window benchmarking can remain only moderately effective under unseen-subject testing, so segment-wise results should not be treated as direct evidence of plug-and-play subject-independent usability.

The present interpretation also aligns with Del Pup *et al.*, who show that subject-based evaluation is more reliable when cross-subject inference is the goal [8]. In that context, the current study contributes a stronger subject-wise reference than segment-level benchmarking because it removes the most obvious subject-overlap pathway, uses a held-out subject in every outer fold, and reports all retained folds transparently.

4.8. Field Significance of the Present Benchmark

The importance of the present result lies not in claiming that 58% accuracy is competitively high in absolute terms, but in showing what a more defensible subject-wise estimate looks like when the full retained cohort is exposed to scrutiny. For future MI-EEG studies, a moderate but transparently derived subject-wise score can be more scientifically valuable than a near-ceiling pooled-window result whose inference target is ambiguous. In that sense, the present paper contributes a calibration point for later architectures, domain-adaptation strategies, compact models, and subject-invariant learning methods.

The benchmark also has practical significance for how claims are framed and reviewed. Segment-level studies can still justify representational promise, proof-of-concept screening, and ablation analysis, but they do not support the same translational claim as subject-wise evaluation. By separating those claim levels explicitly, the present paper reduces the risk that benchmark discrimination is mistaken for plug-and-play readiness and helps align algorithmic reporting with realistic expectations for assistive and neurorehabilitation-oriented BCI development.

4.9. Closest Same-Dataset Subject-Independent Comparisons

Same-dataset subject-independent comparisons are only meaningful when protocol matching is made explicit. Table 5 therefore reports not only headline performance, but also cohort size, validation style, and whether the paper visibly exposes the held-out fold profile. The goal is not a leaderboard; it is a fair description of what each reported number can and cannot support.

Table 5. Closest same-dataset subject-independent or LOSO-like PhysioNet comparisons discussed in the paper. “Fold-level profile shown” indicates whether the paper visibly reports the held-out subject-wise outcome pattern rather than only cohort-level aggregates. One classical CSP-based preprint is included as exploratory context rather than as a definitive benchmark.

Paper	Year	Subjects	Validation style	Reported result	Fold-level profile shown	Fairness note
This work	2026	105	LOSO CV, one held-out subject per fold	58.07% ± 8.27%, (min 38.10%, median 57.14%, max 78.57%)	Yes	Exact retained executable cohort (subjects 1–105) is named, all 105 held-out folds are reported, and cohort-level inference is paired with explicit fold-level variability plots. This is the highest-transparency comparator in the table.
Altamirano CSP+SVM preprint [24]	2026	5	LOSO on a small subject subset	51.11% ± 10.33%	Partial (5 folds)	Useful classical-baseline context, but only five subjects are analysed and the source is a preprint; the comparison is therefore exploratory rather than definitive.
SSDA [25]	2024	105	LOSO	78% ± 3%	No visible full table	Closest protocol match among the checked papers, but the accessible main results emphasise aggregate mean ± standard deviation rather than the complete held-out subject profile.
EEGSym [26]	2022	cohort	LOSO /inter-subject with pretraining and fine-tuning	88.6% ± 9.0%	No visible full table	Strong result, but not directly like-for-like because the pipeline uses a stronger transfer-learning setup and the accessible reporting foregrounds cohort-level inter-subject summaries.
GCNs-Net [6]	2024	109	Group-level cross-subject evaluation	88.57%	No visible full table	High reported PhysioNet performance in a strong IEEE venue, but the accessible description is group-level rather than a strict retained-cohort LOSOCV rerun matched to the present protocol.
EEGNet global validation [20]	2020	105	Subject-based 5-fold global validation	82.43%	No visible full table	Useful subject-based context, but not a strict LOSO study; protocol match is therefore partial rather than full.
Gaussian connectivity-driven EEG imaging [27]	2025	109	Same-dataset deep-learning benchmark; exact LOSO match not explicit in accessible summary	Same-dataset result reported in article	No visible full table	Relevant recent PhysioNet comparator, but task formulation and accessible protocol details are not sufficiently aligned with the present binary retained-cohort LOSOCV study for a like-for-like accuracy claim.
EEG-TriNet++ [28]	2026	PhysioNet cohort	LOSO	70.8% ± 0.9%	No visible full table	Strong recent same-dataset LOSO comparator with repeated-run summaries and macro-F1 reporting, but it still does not expose the complete held-out fold-by-fold subject profile the way the present paper does.

Two points follow from Table 5. First, the present accuracy is plainly below the strongest same-dataset subject-independent reports. That fact should not be hidden. Second, those higher numbers do not answer exactly the same methodological question. Some use transfer learning or stronger pretraining, some use subject-based grouping rather than strict LOSO, and several emphasise only cohort-level mean ± standard deviation instead of exposing the full held-out subject-wise profile.

The exploratory classical row is useful for a narrower reason. Altamirano’s CSP-based preprint reports a best LOSO average accuracy of 51.11% with SVM on a five-subject subset of the same PhysioNet resource. Because the study is both unpublished and cohort-mismatched, it cannot serve as a definitive benchmark. Even so, it suggests that the retained-cohort GAF–PLV result captures

somewhat more cross-subject structure than a simple classical CSP+SVM baseline, while still leaving substantial headroom for robust subject-invariant decoding.

Recent same-dataset studies continue to show this heterogeneity rather than eliminating it. Some recent works remain based on restricted subject subsets or segmentation-based sample construction, whereas others move closer to explicit subject-independent or LOSO-style evaluation [27–29]. This newer literature therefore supports the same interpretive point made throughout the present paper: same-dataset accuracy should be read together with cohort definition, partitioning logic, and fold-level reporting transparency, not in isolation.

The present manuscript is therefore best read as a high-transparency subject-wise benchmark for this representation family. Its distinguishing strengths are explicit retained-cohort definition, complete held-out fold disclosure, and joint reporting of cohort-level and fold-level variability. Relative to the studies in Table 5, those properties make the present comparison set more interpretable even when the raw accuracy is lower. Higher numbers obtained under stronger architectures, partially different task formulations, or less explicit subject-wise reporting do not invalidate the present result; they show that unseen-subject MI-EEG performance depends jointly on model class, training regime, cohort definition, and protocol strength.

4.10. Transparent Cohort Reporting and the Absent Four Subjects

A second methodological issue is the difference between the official dataset size and the retained executable cohort. The official PhysioNet MI resource contains 109 volunteers [9]. However, the available study outputs on which the present paper is based contain fold-level results only for subjects 1–105. We therefore know exactly which four subjects are absent from the available output set: 106, 107, 108, and 109. What we do *not* know from the available files alone is a fully documented machine-readable reason for their absence. This distinction matters. It would be easy to repeat a familiar but unsupported phrase such as “four subjects were excluded due to data inconsistencies” without showing file-level evidence. We chose not to do that.

Instead, the paper adopts a stricter reporting stance. The scientific claim is tied to the 105 retained folds that are actually available in the study outputs, and the absent subjects are stated explicitly rather than hidden. This is more transparent than studies that use reduced cohorts without naming the retained subject IDs or change the effective cohort without showing the fold-level outputs. The same-dataset literature demonstrates how common this ambiguity has become [4,13,15,16]. In that context, exact fold disclosure is part of the core evidence base, not a minor administrative detail.

4.11. Computational Cost

The experiment was run on an NVIDIA GeForce RTX 4090. The average time per epoch was 215.03 s, the average number of trained epochs per fold was 15.70, the approximate runtime per fold was 0.94 h, and the approximate total runtime across 105 folds was 99.05 h. This computational burden helps explain why reduced-cohort benchmarking has persisted historically. More importantly, it shows that even a nearly 100-hour subject-wise rerun on a modern GPU yields only moderate cross-subject performance. That observation underscores why protocol stringency should be considered alongside raw % accuracy.

5. Study Scope and Next Steps

The present study has a clearly defined scope and several next-step priorities. First, the executable rerun package currently supports only the binary task, so the four-class LOSOCV extension should not be inferred from the reported results. Second, although LOSOCV removes the most obvious subject-overlap pathway, the implementation is not nested; nested subject-based selection would provide a stronger estimate when model selection and final testing must be strictly separated [8]. Third, each outer fold uses only a single validation subject, which is statistically fragile. Fourth, strong protocol-matched subject-wise baselines such as CSP+LDA, EEGNet, and other compact MI-EEG references are not yet included in the executable comparison. The exploratory CSP+SVM preprint

cited in this paper is useful only as a small-subset external reference, not as a substitute for a matched baseline rerun. In addition, the absent subjects 106–109 are known from the supplied files, but the reason for their absence is not yet machine-readably documented.

These scope boundaries define the natural next steps: nested subject-wise validation, matched classical and compact-deep baselines, formal ablations, and methods aimed explicitly at subject-invariant learning. Domain adaptation, representation alignment, calibration, and confidence-aware decision rules are especially relevant directions. Accordingly, the present manuscript should be read as a transparent subject-wise benchmark for this representation family rather than as a final estimate of deployable subject-independent performance.

These scope boundaries also motivate a practical reporting standard for future MI-EEG studies. At a minimum, authors should:

1. state the validation unit explicitly (window, trial, subject, or nested subject-based design);
2. report benchmark-oriented discrimination claims and subject-independent transfer claims as separate evidence levels rather than presenting one as proof of the other;
3. report the full subject-wise fold profile rather than only a cohort mean;
4. disclose cohort inclusion and exclusion rules with exact subject identifiers whenever possible;
5. include chance-corrected and/or class-balanced metrics alongside raw accuracy; and
6. provide confidence intervals or equivalent uncertainty estimates for the main performance statistics.

None of these requirements is methodologically extravagant, but together they make it much harder to confuse benchmark discrimination with unseen-subject generalization and much easier to build cumulative, internationally interpretable evidence.

6. Conclusion

Deep learning results in MI-EEG depend not only on model design but also on what the evaluation protocol actually asks the model to do. This paper examined how the scientific meaning of performance changes when our recently proposed GAF–PLV parallel CNN is moved from segment-level benchmarking to subject-wise LOSOCV. On the 105 retained executable subjects, the model achieved $58.07\% \pm 8.27\%$ mean accuracy and $53.48\% \pm 11.19\%$ macro-F1, with substantial subject-to-subject variability.

Read together with the earlier 99.73% segment-level benchmark, these results reveal a clear generalization gap between permissive pooled-window discrimination and unseen-subject transfer. The mean gap is 41.66 points, while the best-, median-, and worst-held-out subject gaps are 21.16, 42.59, and 61.63 points, respectively. The implication is not that segment-wise studies are useless; they remain valuable for rapid representation screening and early method development. The implication is that they should not be interpreted as direct evidence of plug-and-play subject-independent MI-BCI performance.

By making the retained cohort explicit, reporting every available held-out fold, and quantifying uncertainty and heterogeneity rather than hiding them behind a single headline value, this study provides a stronger subject-wise reference for future work on this representation family. The next priorities are nested subject-wise validation, matched classical and compact-deep baselines, and methods designed explicitly for subject-invariant learning. More broadly, the paper’s contribution is to show how a strong proof-of-concept benchmark becomes scientifically more informative and interpretively clearer when examined under subject-wise evaluation. A model that reaches 99.73% under segment-level splits can still fail to provide robust performance for many new users. If MI-EEG research is serious about plug-and-play subject-independent BCI, within-subject discrimination and cross-subject generalization must be reported separately rather than conflated.

Author Contributions: Conceptualization, methodology, software, formal analysis, investigation, data curation, validation, resources, visualization, and writing—original draft preparation, R.L. and H.A.; conceptualization,

problem formulation, validation, supervision, project administration, and writing—review and editing, M.T.S. and W.C.; writing—review and editing and validation, R.N. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The EEG Motor Movement/Imagery dataset analysed in this study is publicly available from PhysioNet at <https://www.physionet.org/content/eegmmidb/1.0.0/>. No new public dataset was created in this study.

Acknowledgments: The authors used ChatGPT (OpenAI) as a language-support tool to assist with drafting, editing, formatting, and improving the clarity of the manuscript. All scientific concepts, methodology, experiments, data analysis, interpretation, and final manuscript decisions were developed, verified, and approved by the authors, who take full responsibility for the content.

Institutional Review Board Statement: Ethical review and approval were waived for this study because it was a secondary analysis of a publicly available, de-identified EEG dataset. No new human participants were recruited, no new biological samples or recordings were collected, and no identifiable personal data were accessed by the authors.

Informed Consent Statement: Patient consent was waived because this study used only publicly available, de-identified data from the PhysioNet EEG Motor Movement/Imagery dataset and did not involve direct interaction with participants or access to identifiable information.

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

References

1. N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "EEG-Based Brain-Computer Interfaces Using Motor-Imagery: Techniques and Challenges," *Sensors*, vol. 19, no. 6, p. 1423, 2019.
2. X. Wang *et al.*, "An in-depth survey on deep learning-based motor imagery electroencephalogram classification," *Artificial Intelligence in Medicine*, vol. 150, Art. no. 102738, 2024.
3. R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.
4. H. Dose, J. S. Møller, H. K. Iversen, and S. Puthusserypady, "An end-to-end deep learning approach to MI-EEG signal classification for BCIs," *Expert Systems with Applications*, vol. 114, pp. 532–542, 2018.
5. W. Huang, W. Chang, G. Yan, Y. Zhang, and Y. Yuan, "Spatio-spectral feature classification combining 3D-convolutional neural networks with long short-term memory for motor movement/imagery," *Engineering Applications of Artificial Intelligence*, vol. 120, 2023.
6. Y. Hou *et al.*, "GCNs-Net: A Graph Convolutional Neural Network Approach for Decoding Time-Resolved EEG Motor Imagery Signals," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7312–7323, 2024.
7. R. Lv, W. Chang, G. Yan, M. T. Sadiq, W. Nie, and L. Zheng, "Enhanced classification of motor imagery EEG signals using spatio-temporal representations," *Information Sciences*, 2025.
8. F. Del Pup, A. Zanolà, L. F. Tshimanga, A. Bertoldo, L. Finos, and M. Atzori, "The role of data partitioning on the performance of EEG-based deep learning models in supervised cross-subject analysis: A preliminary study," *Computers in Biology and Medicine*, vol. 196, Art. no. 110608, 2025.
9. A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, 2000.
10. V. A. Lomelin-Ibarra, A. E. Gutierrez-Rodriguez, and J. A. Cantoral-Ceballos, "Motor Imagery Analysis from Extensive EEG Data Representations Using Convolutional Neural Networks," *Sensors*, vol. 22, no. 16, Art. no. 6093, 2022.
11. K. Roots, Y. Muhammad, and N. Muhammad, "Fusion Convolutional Neural Network for Cross-Subject EEG Motor Imagery Classification," *Computers*, vol. 9, no. 3, Art. no. 72, 2020.
12. R. R. Chowdhury, Y. Muhammad, and U. Adeel, "Enhancing Cross-Subject Motor Imagery Classification in EEG-Based Brain-Computer Interfaces by Using Multi-Branch CNN," *Sensors*, vol. 23, no. 18, Art. no. 7908, 2023.
13. X. Lun, Z. Yu, T. Chen, F. Wang, and Y. Hou, "A Simplified CNN Classification Method for MI-EEG via the Electrode Pairs Signals," *Frontiers in Human Neuroscience*, vol. 14, Art. no. 338, 2020.

14. D. Li, P. Ortega, X. Wei, and A. Faisal, "Model-Agnostic Meta-Learning for EEG Motor Imagery Decoding in Brain-Computer-Interfacing," in *Proc. 10th Int. IEEE/EMBS Conf. Neural Engineering (NER)*, 2021, pp. 527–530.
15. T. Majoros and S. Oniga, "Overview of the EEG-Based Classification of Motor Imagery Activities Using Machine Learning Methods on the PhysioNet Four-Class Motor Imagery Dataset," *Electronics*, vol. 11, no. 15, Art. no. 2293, 2022.
16. H. W. Aung, J. J. Li, Y. An, and S. W. Su, "EEG_GLT-Net: Optimising EEG graphs for real-time motor imagery signals classification," *Biomedical Signal Processing and Control*, vol. 104, Art. no. 107458, 2025.
17. W. Huang, G. Yan, W. Chang, Y. Zhang, and Y. Yuan, "EEG-based classification combining Bayesian convolutional neural networks with recurrence plot for motor movement/imagery," *Pattern Recognition*, vol. 144, Art. no. 109838, 2023.
18. A. Ghimire and K. Sekeroglu, "Classification of EEG Motor Imagery Tasks Utilizing 2D Temporal Patterns with Deep Learning," in *Proc. 2nd Int. Conf. Image Processing and Vision Engineering (IMPROVE)*, 2022, pp. 182–188.
19. W. Huang, W. Chang, G. Yan, Z. Yang, H. Luo, and H. Pei, "EEG-based motor imagery classification using convolutional neural networks with local reparameterization trick," *Expert Systems with Applications*, vol. 187, Art. no. 115968, 2022.
20. X. Wang, M. Hersche, B. Tömekce, B. Kaya, M. Magno, and L. Benini, "An Accurate EEGNet-based Motor-Imagery Brain-Computer Interface for Low-Power Edge Computing," in *Proc. IEEE Int. Symp. Medical Measurements and Applications (MeMeA)*, 2020.
21. J. F. Hwaidi and T. M. Chen, "Classification of Motor Imagery EEG Signals Based on Deep Autoencoder and Convolutional Neural Network Approach," *IEEE Access*, vol. 10, pp. 48071–48081, 2022.
22. C. Fan, B. Yang, X. Li, and P. Zan, "Temporal-frequency-phase feature classification using 3D-convolutional neural networks for motor imagery and movement," *Frontiers in Neuroscience*, vol. 17, Art. no. 1250991, 2023.
23. Z. Wang and T. Oates, "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.
24. J. Altamirano, "Motor Imagery EEG Classification using Common Spatial Patterns and Machine Learning: A Cross-Subject Study," preprint, Mar. 2026.
25. M. Sartipi, M. E. Yaghoubi, and A. M. Nasrabadi, "A subject-independent semi-supervised deep architecture for motor imagery classification from EEG signals," arXiv:2402.09438, 2024.
26. A. Perez-Velasco, E. Santamaria-Vazquez, and V. Martinez-Cagigal, "EEGSym: Overcoming inter-subject variability in motor imagery based BCIs with deep learning," *Journal of Neural Engineering*, vol. 19, no. 5, Art. no. 056018, 2022.
27. A. Gomez-Rivera, D. F. Collazos-Huertas, and others, "Gaussian Connectivity-Driven EEG Imaging for Deep Learning-Based Motor Imagery Classification," *Sensors*, vol. 26, no. 1, Art. no. 227, 2025.
28. A. Tibermacine, I. Naidji, I. E. Tibermacine, L. Mamen, A. Rabehi, and M. Habib, "EEG-TriNet++: A Transformer-Guided Meta-Learning Framework for Robust and Generalizable Motor Imagery Classification," *Bioengineering*, vol. 13, no. 3, Art. no. 307, 2026.
29. X. Lian, C. Liu, C. Gao, and others, "A Multi-Branch Network for Integrating Spatial, Spectral, and Temporal Features in Motor Imagery EEG Classification," *Brain Sciences*, vol. 15, no. 8, Art. no. 877, 2025.

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.