Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

EEG-Conformer for EEG Decoding: A Participant-Wise Evaluation and Ablation Study Across Three Datasets

Ha Rang Cho
Konyang University
Woong Sik Kim
wskim@konyang.ac.kr

Konyang University

Research Article

Keywords: EEG-Conformer, Transformer, Brain–computer interface, Electroencephalography, Ablation study, Participant-wise cross-validation

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11264727/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

EEG-Conformer for EEG Decoding: A Participant-Wise Evaluation and Ablation Study Across Three Datasets

*Ha Rang Cho[1] · Woong Sik Kim[2]**

[1] Department of AI Software Convergence, Konyang University, 158 Gwanjeodong-ro, Seo-gu, Daejeon 35365, Republic of Korea; 26860602@konyang.ac.kr; ORCID: 0009-0001-7645-4012

[2] Department of Artificial Intelligence, Konyang University, 158 Gwanjeodong-ro, Seo-gu, Daejeon 35365, Republic of Korea; wskim@konyang.ac.kr; ORCID: 0000-0002-3916-768X

** Corresponding author: wskim@konyang.ac.kr*

Abstract

Hybrid convolutional–Transformer networks are increasingly used for electroencephalography (EEG) decoding, but their components' contribution and the stability of reported effects are rarely tested. We evaluated the published EEG-Conformer against EEGNet, ShallowConvNet and DeepConvNet on a custom dataset of 48 healthy participants (eyes-open, eyes-closed, hands-up; 19 EEG and 3 accelerometer channels at 500 Hz) and on subsets of PhysioNet and BCI Competition IV-2a, using participant-wise splits. In a single split, EEG-Conformer reached 67.27% three-class accuracy, close to EEGNet (67.00%), and removing the accelerometer channels appeared to raise accuracy by 2.9 points. Five repeats of five-fold cross-validation with different partitions and seeds reversed the accelerometer effect: removing the accelerometer channels lowered accuracy by 1.2 points (95% CI 0.1–2.3) and removing the Transformer by 0.7 points (0.2–1.3), both small against a between-participant SD of 8–9 points. On PhysioNet, the single-split Transformer benefit (2.1 points) disappeared in repeated cross-validation (−0.2 points; 95% CI −1.4 to 1.0). BCI IV-2a leave-one-subject-out validation showed no clear benefit (+0.5 points; 95% CI −1.5 to 2.5). Exploratory tests found no clear loss after downsampling to 125 Hz or restricting the band to 4–13 Hz, whereas keeping only 13–40 Hz cost 5.6 points. Hands-up recall (about 50%) showed no detectable change under band or channel restrictions, with wide intervals. The fixed recording order remains a confound that a within-recording probe could not resolve. Trial-level splitting yielded 13.4-point higher BCI IV-2a accuracy, although training size and test composition also differed; single-split estimates can misstate even the sign of small effects.

Keywords *EEG-Conformer · Transformer · Brain–computer interface · Electroencephalography · Ablation study · Participant-wise cross-validation*

1. Introduction

Brain–computer interfaces (BCIs) that decode scalp electroencephalography (EEG) offer a non-invasive route to restoring communication and control for individuals with severe motor impairment (Abiri et al. 2019; Lotte et al. 2018; Wolpaw and Wolpaw 2012). EEG remains attractive for clinical and translational BCI research owing to its high temporal resolution and comparatively low cost, but scalp recordings suffer from a low signal-to-noise ratio and are highly susceptible to ocular, muscular, and motion artifacts (Craik et al. 2019). Deep learning has substantially advanced EEG decoding (Roy et al. 2019): convolutional neural networks (CNNs) such as EEGNet (Lawhern et al. 2018) and Shallow/DeepConvNet (Schirrmeister et al. 2017) extract local spatial and spectral features efficiently and remain standard baselines, but their fixed, local receptive fields may limit their ability to model relationships between distant channels or long time series (Song et al. 2023). Transformer-based self-attention (Vaswani et al. 2017) offers a complementary mechanism for modeling such long-range dependencies, and hybrid convolutional–Transformer ("Conformer") architectures (Song et al. 2023) have been proposed to combine local and global feature modeling.

A natural hypothesis is that combining CNN-based local feature extraction with Transformer-based global context modeling, and combining EEG with auxiliary sensors such as accelerometers, should improve decoding accuracy and robustness in real-world, non-laboratory recordings. Whether these additions provide a measurable benefit—as opposed to simply adding complexity—is an empirical question best answered by direct ablation under participant-wise evaluation. The present study therefore makes a methodological and empirical contribution rather than an architectural one: we do not propose a new network, but use the previously published EEG-Conformer (Song et al. 2023) as a representative CNN–Transformer model to examine which of its components, and which evaluation choices, determine the accuracies reported for this class of model. The custom recordings were acquired at 500 Hz, higher than the two public benchmarks analyzed here (160 Hz and 250 Hz); whether this finer sampling contributes to decoding is tested directly in Section 3.7.

Specifically, this study contributes: (1) a participant-wise comparison of EEG-Conformer with three widely used CNN baselines (EEGNet, ShallowConvNet and DeepConvNet) on a custom 500 Hz dataset and two public benchmarks; (2) component ablations of the Transformer module and of accelerometer fusion, repeated five times with different participant partitions and initialization seeds and analyzed at the participant level, together with sampling-rate, window-length, band-limit and hands-up sensitivity analyses; (3) a quantification of how evaluation choices change the conclusions, contrasting trial-level with participant-wise splits, a single split with cross-validation, and fold-level with participant-level statistical inference; and (4) a documented, artifact-aware preprocessing pipeline for a proprietary 500 Hz acquisition format, including per-epoch filtering, per-channel artifact rejection and file-level integrity checks.

2. Materials and Methods

2.1. System Overview

EEG-Conformer, originally proposed by Song et al. (2023) as a hybrid convolutional–Transformer architecture for EEG decoding, is used in this study as a representative CNN–Transformer architecture for controlled component evaluation rather than being presented as a newly proposed neural architecture; our implementation adapts the convolutional kernel sizes, pooling, and Transformer configuration (Sections 2.2-2.4) to the channel counts and sampling rates of the three datasets evaluated here, and is not a claim of architectural novelty. EEG-Conformer processes multichannel EEG through four sequential stages: (1) a temporal convolution module, (2) a spatial convolution module, (3) a Transformer-based global context module, and (4) a classification head (Figure 1, Figure 2). The Transformer does not operate on the 2000 input samples. After the temporal and spatial convolutions, average pooling (kernel 75 samples, stride 15 samples) reduces a 4 s epoch at 500 Hz to 129 tokens of dimension 40; at 250 Hz and 160 Hz the same pooling yields 62 and 38 tokens (Table 3). Self-attention therefore acts on a short, strongly smoothed token sequence, and the convolution and pooling kernels, which are fixed in samples, correspond to different physical durations at different sampling rates (Table 3). In this study, "high temporal resolution" refers only to the acquisition rate of the custom dataset (500 Hz). Whether the finer sampling contributes to decoding is examined directly by the resampling, window-length and band-limit analyses (Section 3.7), and the comparison across datasets is interpreted with these sampling-rate-dependent receptive fields in mind.

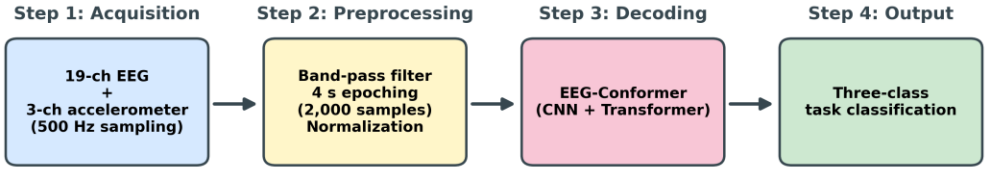


Fig. 1 *Overall data-processing and decoding pipeline, from multimodal 500 Hz EEG acquisition to classification*

2.2. Temporal and Spatial Feature Extraction

Given an input X ∈ R^(C×T) (C channels, T time steps), a temporal convolution (1×25 kernel, 40 output filters, zero-padded to preserve length) is applied along the time axis to extract band-limited temporal features. A subsequent spatial convolution (C×1 kernel, 40→40 filters) collapses the channel dimension, modeling interactions among electrodes (and, where included, the accelerometer channels), analogous in spirit to the depthwise-separable design used in EEGNet (Lawhern et al. 2018), followed by batch normalization, an ELU nonlinearity, average pooling (kernel 1×75, stride 1×15), and dropout (p = 0.5).

2.3. Transformer-Based Global Context Module

Self-attention weights are computed as:

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

where Q, K, and V are the query, key, and value matrices derived from the convolutional feature map and √d_k is a scaling factor (Vaswani et al. 2017). No explicit positional encoding is used in the main model; temporal order is available to the classifier only through the flattened token sequence that follows the encoder. A variant with sinusoidal positional encoding is evaluated in Section 3.3 (+0.54 points, 95% CI −0.17 to +1.25). The encoder consists of a single Transformer layer (d_model = 40, matching the number of convolutional filters; 4 attention heads; feed-forward dimension 160; dropout p = 0.5; GELU activation), applied to the pooled token sequence output by the convolutional stage (129 tokens at 500 Hz).

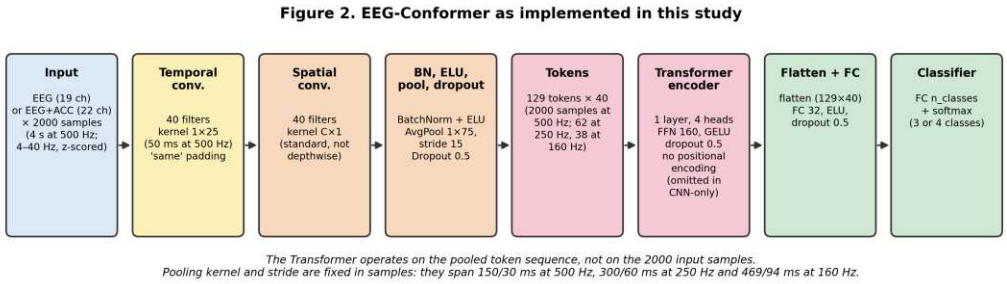


Fig. 2 *Architecture of EEG-Conformer as implemented in this study. The Transformer encoder operates on the pooled token sequence (129 tokens at 500 Hz), has no positional encoding, and is followed by flattening and two fully connected layers*

2.4. Classification and Optimization

Transformer outputs are flattened and passed through a fully connected layer (output dimension 32), ELU activation, dropout (p = 0.5), and a final linear classifier (output dimension equal to the number of classes) with softmax activation. All classifier layers are instantiated at model construction time (before optimizer creation) to guarantee that every trainable parameter is registered with the optimizer — a precaution adopted after an earlier implementation, in which classifier layers were instantiated lazily on the first forward pass, was found to silently exclude the final classification layer from optimization. The network is trained with the cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

using the Adam optimizer (initial learning rate = 0.001, weight decay = 1×10^{-3}, StepLR schedule with step size 30 and decay factor $\gamma = 0.5$), a batch size of 64, and 100 training epochs (a fixed budget selected in advance; no early stopping was used). CNN baselines (EEGNet, ShallowConvNet, DeepConvNet) were trained with the same optimizer settings, batch size, and epoch budget, using their standard published architectures (Lawhern et al. 2018; Schirrmeister et al. 2017) re-implemented for the channel and sample dimensions of each dataset. To support exact reproducibility, all random seeds (NumPy, PyTorch) were fixed and cuDNN was run in deterministic mode (cudnn.deterministic = True, cudnn.benchmark = False).

All hyperparameters (learning rate, weight decay, dropout rate, batch size, and the 100-epoch training budget) were fixed in advance to commonly used default values from the EEGNet and EEG-Conformer literature (Lawhern et al. 2018; Song et al. 2023) and were applied identically to all four models and all three datasets; no hyperparameter search, validation-based model selection, or early-stopping criterion was performed. Because no validation set was held out for tuning, no held-out participant data were used for gradient updates or hyperparameter selection within an evaluation run. The repeated ablation analyses added during revision were motivated by the initial findings and the editorial comments, and the sensitivity analyses are exploratory. This design avoids test-set leakage into hyperparameter selection at the cost of not optimizing per-dataset hyperparameters, which may partly explain the modest accuracy differences among models (Sections 3.1, 3.5, and 3.6).

2.5. Custom EEG Dataset

Data were collected from 50 healthy participants recruited from the university student population (no known neurological or psychiatric conditions; not a clinical or patient sample) using a 19-channel dry-electrode EEG headset (CGX Quick-20r) sampled at 500 Hz, combined with 3 accelerometer channels (22 channels total). Each participant completed three separate ~5-minute recordings corresponding to three states: eyes-open rest, eyes-closed rest, and a sustained hands-up motor task. The three states were recorded in a fixed order for every participant (eyes-closed, then eyes-open, then hands-up); the order was not counterbalanced or randomized. Table 1 lists the electrode-to-region mapping for the 19-channel montage (note that channels P7/P8 correspond to the T5/T6 positions in the classical 10–20 nomenclature). The hands-up task consisted of overt, sustained arm elevation (motor execution rather than motor imagery); electromyography (EMG) was not recorded, and no artifact-correction method targeting muscle activity (e.g., independent component analysis) was applied (Section 2.7). Recordings from this dataset were previously analyzed in a Korean-language study of the effect of data continuity on EEG-based BCI decoding (Kim et al. 2024); the present study uses the same recordings to address a different research question.

Brain Region	Electrodes	Associated Function
Frontal	Fp1, Fp2, F3, F4, Fz, F7, F8	Frontal/prefrontal activity; ocular artifact sensitivity
Central	C3, C4, Cz	Sensorimotor activity related to motor execution
Parietal	P3, P4, Pz	Sensory integration
Temporal / Posterior-Temporal	T3, T4, P7 (T5), P8 (T6)	Temporal lobe activity
Occipital	O1, O2	Visual/occipital alpha response (eyes-open vs. eyes-closed)

Table 1 *Electrode-to-region mapping for the 19-channel EEG montage used in the custom EEG dataset*

2.6. Data Quality Control

Raw recordings were acquired with a 19-channel dry-electrode EEG headset (CGX Quick-20r) and stored in the device's proprietary binary format ("Bioteck Data File," version 3), consisting of a plain-text header (device identifier, sampling rate, and channel names) followed by fixed-size, 32-bit floating-point sample frames; each frame additionally included an embedded per-sample timestamp channel not listed in the header, which was used to automatically detect and truncate any corrupted trailing samples at the end of a recording. During data preparation, an automated content-hash check (MD5 checksum computed over the

full binary content of each parsed recording, i.e., at the level of the entire file rather than individual epochs) was applied to every file to detect accidental file duplication prior to inclusion in the dataset. All 50 participants completed the three recordings, and two were excluded at this stage. Their six recordings were the only ones shorter than 300 s (286.96 s) and ending in an incomplete sample frame. This length was sufficient for 4 s epoching, so the recordings were not unusable because of their duration; rather, the conversion step, most likely because of the incomplete final frame, failed to parse them and extracted no usable epochs from the first of these files and, because the hash was computed on the parsed output, then flagged the other five files as identical to it. A re-audit of the raw files for this revision showed that all 150 raw recordings (50 participants × 3 states) are distinct, including their signal data without the file header, so these exclusions reflect unreadable truncated recordings rather than duplicated data, and the description of them as duplicate recordings in the original submission was incorrect. The final dataset comprises the remaining 48 participants with readable recordings in all three states, and no epoch is duplicated within or across participants (MD5 comparison of all 10,553 epochs).

2.7. Preprocessing Pipeline

Each participant's continuous recording was segmented into non-overlapping 4 s epochs (2000 samples at 500 Hz). To prevent a single large-amplitude artifact from contaminating the entire recording via edge effects of forward–backward (zero-phase) filtering, band-pass filtering (4–40 Hz, 5th-order Butterworth, applied with filtfilt) was performed independently within each 4 s epoch rather than on the full continuous signal; this filter, and the subsequent per-channel Z-score normalization, were applied uniformly to all 22 channels (19 EEG and 3 accelerometer), so that the accelerometer channels entered the network on the same normalized scale as the EEG channels rather than through a separately tuned preprocessing path. Following filtering, an epoch was excluded if more than two of the 19 EEG channels exceeded a ±500 μV amplitude threshold, a criterion chosen to tolerate chronically noisy individual channels (observed in some recordings, e.g., due to imperfect dry-electrode contact) without discarding otherwise usable epochs; the accelerometer channels were not included in this rejection criterion. Retained epochs were then Z-score normalized per channel. Figure 3 shows representative filtered signals from three EEG channels and one accelerometer channel for a single participant.

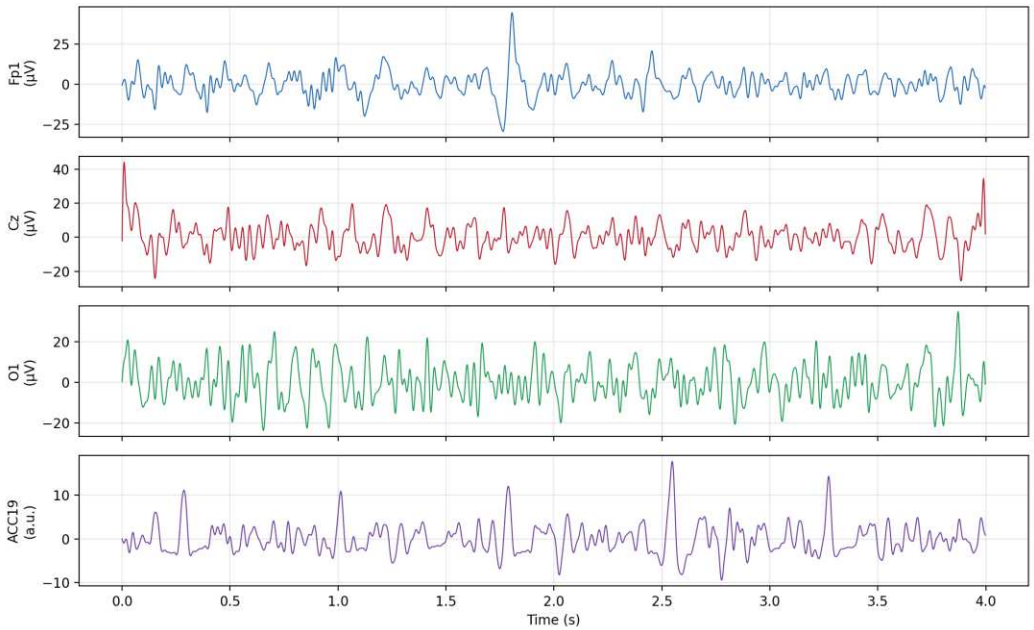


Fig. 3 *Representative 4 s segment of band-pass-filtered (4–40 Hz) EEG (Fp1, Cz, O1) and accelerometer (ACC19) signals from one participant during the eyes-open condition, illustrating typical post-preprocessing amplitude and morphology*

This per-epoch filtering and artifact-rejection strategy was adopted after an initial diagnostic step revealed that applying the band-pass filter to an entire ~5-minute recording allowed an isolated, large-amplitude artifact to propagate—via the two-pass filter's impulse response—into otherwise clean portions of the recording several seconds away. Segmenting before filtering confined this effect to the epoch in which the artifact occurred.

2.8. Experimental Setup

Item	Specification
GPU	NVIDIA GeForce RTX 3080 (10 GB)
Software	Python 3.11.11, PyTorch 2.3.0 (CUDA 11.8), NumPy 2.4.6, SciPy 1.17.1, scikit-learn 1.8.0 (revision experiments)
Sampling rate	500 Hz
Epoch length	4 s (2000 samples)
Channels	19 EEG + 3 accelerometer (22 total)

Table 2 *Hardware and data specification*

Final dataset: 10,553 epochs from 48 participants (class distribution: eyes-open = 3522, eyes-closed = 3510, hands-up = 3521). Participants, rather than epochs, were split into training (38 participants, 8344 epochs) and test (10 participants, 2209 epochs) sets, so that no participant contributed epochs to both sets.

Because the three datasets evaluated in this study (Sections 3.4-3.6) differ substantially in sampling rate, channel count, epoch length, and number of classes, Table 3 summarizes these specifications alongside the participant-wise train/test split used for each. The same four model architectures (EEGNet, ShallowConvNet, DeepConvNet, EEG-Conformer) were applied to all three datasets; only the input dimensions (channel count and number of time samples) and the final classification layer's output size were adapted per dataset, with all other architectural hyperparameters (Sections 2.2-2.4) and training settings (Section 2.4) held fixed.

Dataset	Sampling Rate	Channels	Epoch Samples	Classes	Train/Test Participants	Tokens	Pooling kernel / stride
Custom (Section 2.5)	500 Hz	22 (19 EEG + 3 ACC)	2000	3	38 / 10	129	150 ms / 30 ms
PhysioNet (Section 3.5)	160 Hz	64	640	3 (EO/EC/hands)	39 / 10	38	469 ms / 94 ms
BCI Competition IV-2a (Section 3.6)	250 Hz	22	1001	4	7 / 2	62	300 ms / 60 ms

Table 3 *Summary of the three datasets used for evaluation, including sampling rate, channel count, epoch length, number of classes, and the participant-wise train/test split*

2.9. Statistical Analysis

All statistical tests used a significance threshold of alpha = 0.05 and were implemented in SciPy (scipy.stats). Two categories of comparison are reported. (1) Model comparisons under cross-validation (Sections 3.1 and 3.6). Because EEG-Conformer and the comparator model (EEGNet) were trained and evaluated on identical splits, their accuracies are paired; pairing alone does not remove dependence induced by overlapping training sets. The two levels at which this comparison is reported answer different questions. The fold-level paired t-test uses one paired observation per fold (n = 5 for the custom dataset); its unit is the trained model, so it reflects variability across training partitions, but with four degrees of freedom it has little power and is sensitive to any single fold. The participant-level analysis uses one paired observation per held-out participant (n = 48), which is possible because each participant was tested exactly once per model; its unit is the participant, so it addresses generalization to new individuals, and it was evaluated with both a paired t-test and a Wilcoxon signed-rank test to guard against non-normal per-participant accuracies.

The participant-level observations are, however, not fully independent: participants in the same fold were classified by the same trained model, so fold-specific training effects are shared within each group of about ten participants. Because neither test models this nesting, neither analysis fully accounts for dependence induced by shared models and overlapping training sets; we therefore report both and base our conclusions on their joint pattern rather than on either test alone. Under leave-one-subject-out cross-validation on BCI Competition IV-2a, each fold contains exactly one test participant, so the two levels coincide (n = 9). (2) Accelerometer signal characteristics (Section 3.3): to test whether raw accelerometer variability differed across the three task conditions, we first computed a one-way ANOVA treating every epoch as an independent observation; because many epochs originate from the same participant, this epoch-level test is reported as an exploratory analysis only, since it is subject to pseudo-replication and likely overstates the effective sample size and statistical significance. The primary participant-level test is a one-way repeated-measures ANOVA computed after averaging accelerometer variability within each participant for each condition (n = 48 participants, 3 within-participant conditions), which avoids pseudo-replication and is the basis for the conclusions drawn in Sections 3.3 and 4 regarding accelerometer informativeness.

Repeated cross-validation and factorial analysis. To estimate component effects independently of a single partition, each configuration was evaluated with five-fold participant-wise cross-validation repeated for five seeds. Each seed used a different random assignment of participants to folds (identical for all configurations) and a different weight initialization, so that every participant contributed one test accuracy per seed and configuration (25 runs per configuration); partition and initialization were therefore changed together rather than crossed, and their contributions to variability cannot be separated. The two ablated factors, accelerometer channels and Transformer, form a 2 × 2 design. For each participant and seed we computed the accelerometer effect (the average of the EEG-only minus full contrasts over the two Transformer conditions), the Transformer effect (the average over the two accelerometer conditions) and their interaction, averaged these over seeds within participants, and tested the mean over the 48 participants with a t-test (47 degrees of freedom). Averaging the contrasts over seeds within each participant accounts for the repeated observations of the same participant, but repetition does not remove the dependence among participants evaluated by the same fold-specific model or trained on overlapping sets; no mixed-effects model was fitted, so the confidence intervals and p values are approximate. Variance components (between participants, and between repeats within a participant) came from a one-way random-effects analysis of variance. Holm–Bonferroni correction was applied across the three factorial tests and, separately, across the four pairwise contrasts; the accompanying 95% confidence intervals are unadjusted and should not be read as simultaneous intervals. The sensitivity analyses (Section 2.10) compared each variant with the reference model trained on the same partition, using paired participant-level t-tests and Wilcoxon tests with Holm correction within each analysis family; these analyses are exploratory.

2.10. Sensitivity Analyses and Order-Effect Probe

The sensitivity analyses used the EEG-only EEG-Conformer (convolutional front end and Transformer; 19 EEG channels) as reference, with each variant trained on the same participant partition. Sampling rate: epochs were decimated from 500 Hz to 250 and 125 Hz with anti-aliasing (the data had already been band-limited to 4–40 Hz); the network was either left unchanged, so that kernels fixed in samples span longer durations, or its temporal-convolution and pooling kernels and stride were scaled to span the same durations as at 500 Hz (kernel 12 or 6 samples, pooling kernel 38 or 19 samples, stride 8 or 4 samples at 250 or 125 Hz). Band limit: a zero-phase fourth-order Butterworth band-pass filter was applied within the existing 4–40 Hz band (4–30, 4–20, 4–13 and 13–40 Hz), followed by per-channel Z-score normalization. Window length: each 4 s epoch was divided into two 2 s or four 1 s non-overlapping windows that inherited the participant and label, and test accuracy was computed on windows. Hands-up sensitivity: band limits of 4–20 and 4–13 Hz, exclusion of frontal and temporal channels (Fp1, Fp2, F7, F8, F3, Fz, F4, T3, T4, P7, P8, leaving C3, Cz, C4, Pz, P3, P4, O1, O2), and both restrictions combined. Order-effect probe: because the states were recorded in a fixed order, we estimated how much time-related signal is available within a recording. Each recording was divided at its midpoint into a first and a second half, and a linear discriminant

classifier on log band power (theta, alpha, beta and gamma at 19 channels) was trained to separate the halves in five-fold participant-wise cross-validation; the balanced accuracy in held-out participants (chance 50%) was tested against 0.5 for each state with a t-test and a Wilcoxon test. This probe measures time-related signal within recordings, not differences between blocks. All sensitivity variants were evaluated by five-fold participant-wise cross-validation on the same single partition and initialization seed (seed 0). No equivalence margin was prespecified, and failure to reject a difference is not evidence of equivalent performance.

2.11. Use of Generative Artificial Intelligence

Generative artificial intelligence (Claude, Anthropic) was used to write the analysis code and to draft and edit the manuscript and the response to the editor under the authors' direction. The authors designed the study, collected the data, ran all experiments, verified every reported result against the output files, and take full responsibility for the content of this article.

3. Results

3.1. Model Comparison

EEG-Conformer was compared against three CNN baselines under identical preprocessing, data split, and training configuration (Table 4).

Model	Test Accuracy (%)
DeepConvNet (Schirrmeister et al. 2017)	64.83
ShallowConvNet (Schirrmeister et al. 2017)	64.69
EEGNet (Lawhern et al. 2018)	67.00
EEG-Conformer	67.27

Table 4 *Three-class (eyes-open / eyes-closed / hands-up) test accuracy on the custom EEG dataset (chance level = 33.33%)*

All four models substantially exceeded chance level, suggesting that task-relevant information was extracted from the EEG. On this single 38/10 participant split, EEG-Conformer obtained the highest pooled accuracy, but the margin over EEGNet was small (0.27 percentage points); a single split cannot establish whether this difference is representative, motivating the participant-wise cross-validation reported next.

To assess whether the small margin observed on the single split was representative, we additionally performed 5-fold participant-wise cross-validation (participants grouped into 5 folds of 9-10 participants each, with all folds evaluated using the identical fixed hyperparameters and training procedure of Section 2.4) for EEG-Conformer and EEGNet, the two models separated by the smallest margin on the single split. Table 5 reports the resulting per-fold accuracies.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean ± SD
EEG-Conformer	67.68	66.82	64.97	59.77	64.18	64.68 ± 3.08
EEGNet	65.87	55.94	64.11	59.02	63.21	61.63 ± 4.05

Table 5 *Five-fold participant-wise cross-validation accuracy (%) for EEG-Conformer and EEGNet on the custom EEG dataset*

Across the five folds, EEG-Conformer outperformed EEGNet in every fold (per-fold differences: +1.81, +10.88, +0.86, +0.75, and +0.97 percentage points; mean difference +3.05 pp), a larger and more consistent margin than the 0.27-percentage-point difference observed on the original single split. A paired t-test on the five fold-wise differences alone was not statistically significant ($t = 1.55$, $p = 0.195$, $df = 4$), reflecting the limited statistical power of only five paired observations. Because each of the 48 participants was tested in exactly one fold for each model, we additionally analyzed accuracy at the participant level ($n = 48$),

which provides substantially greater statistical power than the fold-level comparison. Mean per-participant accuracy was 64.81% +/- 10.59% for EEG-Conformer and 61.67% +/- 11.02% for EEGNet (mean difference +3.15 pp); this difference was nominally significant by both a paired t-test ($t = 2.24$, $p = 0.030$) and a Wilcoxon signed-rank test ($p = 0.014$). The fold-level mean difference was strongly influenced by a single fold (+10.88 pp); in the other four folds, EEG-Conformer led by 0.75 to 1.81 pp. Given this and the dependence of participant-level observations within folds (Section 2.9), we interpret the result as a directionally consistent difference in this partition, with nominal participant-level evidence but no dependence-adjusted demonstration of an advantage; it should not be read as robust architectural superiority. The cross-validated means (64.68-64.81% and 61.63-61.67%, fold- and participant-level respectively) are also several percentage points lower than the corresponding single-split accuracies (67.27% and 67.00%), indicating that the original 38/10 split was a relatively favorable partition for both models and that cross-validated estimates provide a more representative picture of expected performance on unseen participants.

Because the reported 67.27% figure is a single pooled accuracy across 2209 epochs from only 10 held-out participants, we additionally computed accuracy separately for each of the 10 test participants to characterize between-participant variability. Per-participant accuracy for EEG-Conformer ranged from 44.59% to 91.89% (mean ± SD: 67.27% ± 14.11%; 95% CI, computed via a t-distribution with 9 degrees of freedom: [57.17%, 77.37%]). This substantial spread indicates considerable inter-participant variability in decoding accuracy and uncertainty in the participant-mean estimate; the participant-mean interval should not be interpreted as a confidence interval for epoch-pooled accuracy, and evaluation on a larger and/or cross-validated set of held-out participants would be needed to narrow this interval.

3.2. Per-Class Performance

Class	Precision	Recall	F1-score
Eyes Open	0.5645	0.4103	0.4752
Eyes Closed	0.9531	0.9128	0.9325
Hands Up	0.5294	0.6955	0.6012
Macro average	0.6823	0.6729	0.6696

Table 6 *Per-class precision, recall, and F1-score for EEG-Conformer on the held-out test set (n = 2209 epochs from 10 participants)*

True \ Predicted	Eyes Open	Eyes Closed	Hands Up
Eyes Open (n=736)	302	20	414
Eyes Closed (n=734)	21	670	43
Hands Up (n=739)	212	13	514

Table 7 *Confusion matrix for EEG-Conformer on the held-out test set (rows = true class, columns = predicted class)*

Eyes-closed epochs were decoded with the highest reliability (F1 = 0.93), consistent with the well-documented increase in occipital alpha-band (8–13 Hz) power that occurs when the eyes close (Berger 1929). In contrast, eyes-open and hands-up epochs were frequently confused in both directions (414 eyes-open epochs misclassified as hands-up; 212 hands-up epochs misclassified as eyes-open), while confusions with eyes-closed were rare for either class. This pattern suggests that the model relied substantially on the visual/occipital state (eyes open vs. closed) as a discriminative feature, and had comparatively less success isolating the motor-execution component of the hands-up condition from the shared "eyes-open" state. Because the hands-up condition involved overt arm elevation, part of what distinguished it from the eyes-open condition may have been muscle or movement artifact rather than sensorimotor cortical activity (Section 4).

3.3. Ablation Study

To isolate the contribution of the accelerometer channels and the Transformer module, four configurations were trained and evaluated under the same procedure: the full model (EEG plus accelerometer, convolutional front end plus Transformer), the EEG-only model, and the two corresponding CNN-only models without the Transformer. Table 8 reports the original single 38/10 split together with five-fold participant-wise cross-validation repeated five times, each with a different participant partition and initialization seed (Section 2.9).

Configuration	Single split (%)	Δ single split (pp)	Repeated CV (%)	Repeat SD (pp)	Δ repeated CV (pp)
Full model (EEG + ACC, CNN + Transformer)	67.27	—	64.16	0.43	—
EEG-only, CNN + Transformer	70.21	+2.94	62.91	0.63	−1.24
EEG + ACC, CNN-only (no Transformer)	67.22	−0.05	63.37	0.31	−0.79
EEG-only, CNN-only	66.86	−0.41	62.22	0.60	−1.94

Table 8 *Ablation results on the custom EEG dataset: original single 38/10 split (accuracy pooled over test epochs) and five-fold participant-wise cross-validation repeated five times with different participant partitions and initialization seeds (accuracy averaged over the 48 participants; repeat SD is the SD of the participant-mean accuracy across the five paired partition/initialization repeats; Δ is the difference from the full model in the repeated analysis)*

In the single split, three configurations lay within 0.41 points of each other (66.86–67.27%), whereas the EEG-only model with the Transformer reached 70.21%. Relative to the full model this appeared as a 2.94-point benefit of removing the accelerometer channels and, within the EEG-only models, as a 3.35-point benefit of the Transformer. The repeated analysis reversed the first effect and removed the second (Table 8). Averaged over the two Transformer conditions, removing the accelerometer channels lowered accuracy by 1.20 points (95% CI −2.27 to −0.12; $t(47) = -2.23$, $p = 0.030$), and averaged over the two accelerometer conditions, removing the Transformer lowered it by 0.74 points (−1.29 to −0.20; $t(47) = -2.73$, $p = 0.009$); the interaction estimate was close to zero (−0.09 points, −0.73 to +0.54; $p = 0.768$). With Holm correction across these three tests the Transformer effect remained significant ($p = 0.027$) and the accelerometer effect did not ($p = 0.060$). Pairwise contrasts gave the same picture; after Holm correction over the four contrasts only the removal of the Transformer in the presence of the accelerometer channels was significant (−0.79 points, 95% CI −1.32 to −0.26; Holm $p = 0.018$; 35 of 48 participants were more accurate with the Transformer). Both estimated effects are small compared with the variation between participants (SD 8.3–9.3 points) and with the variation of a participant's accuracy across partitions and seeds (SD 3.9–4.3 points), and the single-split estimate of the accelerometer effect lay well outside the repeated-analysis estimate, showing that one partition can misstate even the sign of a small effect. Adding sinusoidal positional encoding to the EEG-only model changed accuracy by +0.54 points (95% CI −0.17 to +1.25; $p = 0.131$). A post hoc analysis of the raw accelerometer amplitude indicated that these channels carry task-related information. The participant-level repeated-measures ANOVA on accelerometer standard deviation, the primary test ($n = 48$), was significant ($F(2,94) = 6.33$, $p = 0.0026$; participant means: eyes-open 2.84, eyes-closed 2.73, hands-up 2.99), as was the exploratory epoch-level ANOVA ($F = 49.24$, $p = 5.2e\text{-}22$; $n = 10{,}553$ epochs), which treats epochs from the same participant as independent and does not provide independent confirmatory evidence. The small benefit of including the accelerometer channels in the repeated analysis is consistent with this information, but its size, about one point, does not justify a stronger conclusion, and we did not test alternative normalization or fusion schemes.

3.4. Exploratory Model Interpretation via Grad-CAM

To investigate why eyes-open and hands-up epochs were frequently confused (Section 3.2), we conducted an exploratory interpretability analysis; because it examines only the first convolutional layer (see caveat below), it should be treated as a preliminary probe of the model's early spatial representations rather than

definitive evidence about the full network's decision process. Grad-CAM (Selvaraju et al. 2017) was applied to the temporal-convolution layer of the trained EEG-Conformer model (the first convolutional layer, prior to the depthwise spatial convolution that collapses the electrode dimension), yielding a per-channel importance score for each classified epoch. Importance scores (normalized to [0, 1]) were averaged over 30 epochs per group for five groups: correctly classified eyes-open, eyes-open misclassified as hands-up, correctly classified hands-up, hands-up misclassified as eyes-open, and correctly classified eyes-closed. For each group, the mean importance of the three sensorimotor channels (C3, Cz, C4) was compared against the mean importance of the remaining 16 EEG channels.

Across all five groups, the ratio of sensorimotor to non-sensorimotor channel importance was close to 1.0 (range: 0.93 to 0.99; eyes-open correct: 0.93; eyes-open misclassified as hands-up: 0.96; hands-up correct: 0.99; hands-up misclassified as eyes-open: 0.98; eyes-closed correct: 0.97). At this first convolutional layer, no marked preferential importance of sensorimotor channels was observed for any condition, including the hands-up motor task; because this analysis reflects only the earliest stage of spatial processing (see caveat below), it should be read as showing an absence of early-layer sensorimotor emphasis rather than as evidence that the full network disregards sensorimotor information entirely. This observation is consistent with the interpretation proposed in Section 4: rather than concentrating on channels over the sensorimotor cortex, EEG-Conformer's early representations were roughly evenly distributed across the scalp, which may contribute to the observed difficulty distinguishing an eyes-open resting state from a motor-execution task that, in this montage and paradigm, may not have produced a sufficiently localized or strong sensorimotor signal (e.g., a clear mu-rhythm decrease over C3/C4) relative to broader, non-specific activity shared with the eyes-open condition. A methodological caveat applies to this analysis: because Grad-CAM was computed at the first convolutional layer, before cross-channel information is combined by the subsequent spatial convolution, the resulting channel-importance estimates reflect only the earliest stage of spatial processing and may not fully capture spatial patterns that emerge deeper in the network.

3.5. Comparison with a Public Benchmark (PhysioNet)

As a further check of generalization beyond the custom EEG dataset, EEG-Conformer and the three CNN baselines were trained and evaluated on a 3-class task built from the PhysioNet EEG Motor Movement/Imagery Dataset (Goldberger et al. 2000): the eyes-open resting baseline (run 1), the eyes-closed resting baseline (run 2), and executed both-fists movement (run 5, T1 events). The baseline runs were cut into non-overlapping 4 s epochs from their first 60 s (15 epochs per run) and the movement events into 4 s epochs from event onset (64 channels, 160 Hz; 1832 epochs from 49 participants: 735 eyes-open, 735 eyes-closed and 362 hands). Participants S001–S050 of the 109-participant dataset were requested; S014 was excluded because its eyes-open baseline recording lasted 59.99 s, shorter than the 60 s window used for epoching, and one hands event of participant S034 that extended beyond the end of the recording was dropped. As in the custom dataset, the three states come from separate runs, so the confound between state and recording block described in the limitations also applies here. The same architectures, training procedure, and per-epoch preprocessing (band-pass filtering and Z-score normalization) as for the custom dataset were used. Following a root-cause review of the extraction pipeline, an off-by-one sample-count mismatch present in an earlier version of this analysis was traced to an MNE epoch-construction step that silently dropped a small number of trials extending past the end of a recording without correspondingly adjusting the label count; the extraction was corrected to derive labels directly from the number of retained epochs, and participant identifiers were additionally recovered, enabling a participant-wise (rather than trial-level) split (39 participants for training, 10 for testing; 1457/375 epochs).

Model	Test Accuracy (%)
DeepConvNet (Schirrmeister et al. 2017)	72.53
EEGNet (Lawhern et al. 2018)	79.73
EEG-Conformer	81.07
ShallowConvNet (Schirrmeister et al. 2017)	82.40

Table 9 *Three-class (eyes-open / eyes-closed / hands) test accuracy on a PhysioNet subset (participant-wise split, 39 train / 10 test participants; chance level = 33.33%)*

True \ Predicted	Eyes Open	Eyes Closed	Hands
Eyes Open (n=150)	115	11	24
Eyes Closed (n=150)	13	135	2
Hands (n=75)	17	4	54

Table 10 *Confusion matrix for EEG-Conformer on the PhysioNet subset (participant-wise split; rows = true class, columns = predicted class)*

Under the participant-wise split, ShallowConvNet (82.40%) slightly outperformed EEG-Conformer (81.07%), and all models substantially exceeded both chance level and the accuracy obtained on the custom EEG dataset, although this should not be read as evidence of better data quality: the three states again come from separate runs, and the eyes-closed state is identified with high recall (90%), as in the custom dataset. In the single split, removing the Transformer module reduced accuracy by 2.14 percentage points (81.07% → 78.93%; Table 11). Repeating the evaluation as five-fold participant-wise cross-validation repeated five times, each with a different participant partition and initialization seed (49 participants; EEG-Conformer 74.04%, CNN-only variant 74.25%, EEGNet 72.99%) removed this effect: EEG-Conformer minus CNN-only was −0.21 points (95% CI −1.40 to +0.98; paired t-test p = 0.721, Wilcoxon p = 0.975), and EEG-Conformer minus EEGNet was +1.04 points (−1.12 to +3.20; p = 0.338 and 0.163). The 2.14-point effect of the single split therefore did not generalize across partitions, and the lower repeated-cross-validation accuracy (74.0% versus 81.1%) is another instance of the split dependence examined throughout this study.

Configuration	Single split (%)	Repeated CV (%)	Repeat SD (pp)
EEG-Conformer (CNN + Transformer)	81.07	74.04	0.87
CNN-only (no Transformer)	78.93	74.25	0.91
EEGNet	79.73	72.99	1.12

Table 11 *Transformer ablation and EEGNet comparison on the PhysioNet subset: single participant-wise split (accuracy pooled over test epochs) and five-fold participant-wise cross-validation repeated five times with different participant partitions and initialization seeds (accuracy averaged over the 49 participants; repeat SD is the SD of the participant-mean accuracy across the five paired partition/initialization repeats)*

3.6. Comparison with a Second Public Benchmark (BCI Competition IV-2a)

The same models and procedure were additionally applied to the 4-class BCI Competition IV-2a dataset (Brunner et al. 2008) (left hand, right hand, feet, tongue motor imagery; 22 EEG channels, 250 Hz, 1001 samples per trial), using the full standard cohort of 9 participants defined by the dataset (subjects A01-A09) and both recorded sessions (training and evaluation) per participant, retrieved via the MOABB toolbox's default motor-imagery paradigm, which pools both sessions. This yields 9 participants x 2 sessions x 4 classes x 72 trials per class = 5184 trials in total, consistent with the dataset's official design; no participants or trials were excluded. The MOABB default motor-imagery paradigm band-pass filters the signals at 8–32 Hz and extracts the 4 s motor-imagery interval (1001 samples at 250 Hz); per-epoch Z-score normalization was applied without further filtering, so the effective analysis band for this dataset (8–32 Hz) differs from that of the custom dataset (4–40 Hz). Following a root-cause review of the extraction pipeline, participant identifiers were recovered directly from the metadata returned by the MOABB data-loading paradigm (which had previously been discarded rather than retained), enabling a proper participant-wise (rather than trial-level) train/test split (7 participants for training, 2 for testing; 4032/1152 epochs).

An initial trial-level (80%/20%) split of the same dataset, reported in an earlier version of this analysis, had produced a substantially higher EEG-Conformer accuracy of 59.40%. With participant-wise splitting, accuracy dropped to 46.01%—a 13.4-percentage-point difference consistent with the data leakage inherent in trial-level splitting (epochs from the same participant appearing in both training and test sets). This

comparison, retained here for transparency, illustrates concretely why participant-wise evaluation is necessary for EEG decoding studies; the PhysioNet comparison in Section 3.5 was re-evaluated with the same participant-wise approach for this reason.

Model	Test Accuracy (%)
DeepConvNet (Schirrmeister et al. 2017)	40.19
ShallowConvNet (Schirrmeister et al. 2017)	45.14
EEG-Conformer	46.01
EEGNet (Lawhern et al. 2018)	49.31

Table 12 *Four-class (left hand / right hand / feet / tongue) test accuracy on BCI Competition IV-2a (participant-wise split, 7 train / 2 test participants; chance level = 25%)*

True \ Predicted	Class 0	Class 1	Class 2	Class 3
Class 0 (n=288)	77	50	126	35
Class 1 (n=288)	31	147	87	23
Class 2 (n=288)	30	25	215	18
Class 3 (n=288)	29	47	121	91

Table 13 *Confusion matrix for EEG-Conformer on BCI Competition IV-2a (participant-wise split; rows = true class, columns = predicted class)*

Under the participant-wise split, EEGNet (49.31%) outperformed EEG-Conformer (46.01%), reversing the ranking observed under the trial-level split. Class 2 showed markedly higher recall (74.7%) but lower precision (39.2%) than the other three classes, indicating frequent prediction of Class 2, which increased its recall while reducing its precision and the recall of other classes. In the single split, removing the Transformer module did not reduce accuracy (46.01% → 46.70%, +0.69 pp; Table 14), as had also been the case on this dataset under the trial-level split; the leave-one-subject-out estimate is reported below.

Configuration	Accuracy (%)	Δ vs. Full Model (pp)
Full model (CNN + Transformer)	46.01	—
CNN-only (no Transformer)	46.70	+0.69
Full model, leave-one-subject-out (5 seeds)	44.24	—
CNN-only, leave-one-subject-out (5 seeds)	43.72	−0.52 (−2.51 to +1.48)
EEGNet, leave-one-subject-out (5 seeds)	44.40	+0.16 (−1.34 to +1.66)

Table 14 *Transformer ablation and EEGNet comparison on BCI Competition IV-2a (participant-wise split; first two rows: single 7/2 split, accuracy pooled over test epochs; last three rows: leave-one-subject-out cross-validation, mean over five seeds and nine participants, with the 95% CI of the difference across participants)*

As a further, more rigorous check on this dataset, we additionally performed full 9-fold leave-one-subject-out (LOSO) cross-validation for EEG-Conformer and EEGNet, using every one of the 9 available participants as the held-out test set exactly once (rather than the single 7/2 split reported above). Under full LOSO, mean accuracy was 44.02% +/- 16.17% for EEG-Conformer and 44.14% +/- 18.06% for EEGNet (mean difference -0.12 pp), and a paired t-test across the 9 subject-level observations found no significant difference between the two models ($t = -0.12$, $p = 0.911$). This LOSO-based estimate is lower than both the trial-level (59.40%) and the single 7/2-split (46.01% and 49.31% for EEG-Conformer and EEGNet, respectively) estimates reported above, reinforcing the pattern that more rigorous, fully participant-wise validation yields more conservative and more representative accuracy estimates; it also indicates that, when averaged over all 9 possible held-out participants rather than a single arbitrary pair, no significant difference between EEG-Conformer and EEGNet was found on this dataset, in contrast to the custom EEG dataset, where EEG-Conformer's advantage was significant across participants but not across folds (Section 3.1).

Considered together, the Transformer's contribution was small and not consistent across evaluations. Single-split participant-wise estimates of the Transformer's contribution ranged from −0.7 to +3.4 points across the datasets (Tables 8, 11 and 14), whereas the repeated-cross-validation estimates were +0.7 to +0.8

points on the custom dataset (Section 3.3) and −0.2 points on PhysioNet (Section 3.5); the corresponding leave-one-subject-out estimate on BCI Competition IV-2a follows below.

To place the Transformer ablation on the same footing, we repeated it on BCI Competition IV-2a with full leave-one-subject-out cross-validation over five seeds (9 folds × 5 seeds × 3 configurations; Table 14). EEG-Conformer reached 44.24%, the CNN-only variant 43.72% and EEGNet 44.40% (nine participants, chance 25%); EEG-Conformer minus CNN-only was +0.52 points (95% CI −1.48 to +2.51; paired t-test p = 0.566, Wilcoxon p = 0.910; the smallest attainable two-sided Wilcoxon p with nine participants is 0.0039), and EEG-Conformer was higher for only three of the nine participants. EEG-Conformer minus EEGNet was −0.16 points (95% CI −1.66 to +1.34; paired t-test p = 0.810, Wilcoxon p = 0.652), consistent with the single-seed leave-one-subject-out comparison above; Holm-corrected p values were 1.000 for both contrasts. Accuracy varied far more between participants (SD 14.4–16.3 points) than between seeds (SD 0.4–0.9 points).

3.7. Sensitivity Analyses on the Custom Dataset

We examined whether the 500 Hz acquisition rate, the analysis band and the window length contributed to decoding, using the EEG-only EEG-Conformer trained on one participant partition as reference (accuracy 63.02%; recall eyes-open 48.2%, eyes-closed 88.9%, hands-up 52.0%). Exploratory comparisons did not detect a clear accuracy change after reducing the sampling rate from 500 to 250 or 125 Hz (250 Hz with kernels fixed in samples: +0.07 points, 95% CI −1.15 to +1.28; 250 Hz with kernels scaled to the same duration: −0.84, −2.04 to +0.36; 125 Hz: −0.06, −1.51 to +1.39, and −0.07, −0.99 to +0.84), although these intervals are conditional on one partition and initialization and are not equivalence tests (Table 15). Limiting the band to 4–30, 4–20 or 4–13 Hz did not produce a statistically detectable accuracy change (−0.26, +0.07 and −1.15 points; none significant), whereas keeping only 13–40 Hz lowered it by 5.59 points (95% CI −8.05 to −3.14; Holm-corrected p < 0.001), almost entirely through eyes-closed recall (−15.29 points, −20.77 to −9.81). Shorter windows reduced accuracy: −0.71 points for 2 s (−1.69 to +0.28) and −2.30 for 1 s (−3.79 to −0.81; Holm-corrected p = 0.007). These results are consistent with a substantial contribution of activity below 13 Hz, including the alpha increase with eye closure, rather than of fine temporal detail.

Condition	Δ accuracy (pp)	95% CI	Holm p (t / W)
250 Hz, kernels fixed in samples	+0.07	−1.15 to +1.28	1.000 / 1.000
250 Hz, kernels scaled to equal duration	−0.84	−2.04 to +0.36	0.666 / 0.600
125 Hz, kernels fixed in samples	−0.06	−1.51 to +1.39	1.000 / 1.000
125 Hz, kernels scaled to equal duration	−0.07	−0.99 to +0.84	1.000 / 1.000
Band 4–30 Hz	−0.26	−1.34 to +0.82	1.000 / 1.000
Band 4–20 Hz	+0.07	−1.58 to +1.72	1.000 / 1.000
Band 4–13 Hz	−1.15	−3.17 to +0.87	0.772 / 0.784
Band 13–40 Hz	−5.59	−8.05 to −3.14	<0.001 / <0.001
Window 2 s	−0.71	−1.69 to +0.28	0.155 / 0.161
Window 1 s	−2.30	−3.79 to −0.81	0.007 / 0.018

Table 15 *Sampling-rate, band-limit and window-length sensitivity of the EEG-only EEG-Conformer (one participant partition and seed). Differences from the reference configuration (accuracy 63.02%) in participant-level accuracy with 95% confidence intervals and Holm-corrected p values (paired t-test / Wilcoxon test) within each analysis family*

Hands-up recall was 52.0% in the reference model, and hands-up epochs were classified as eyes-open in 40.6% of cases (Table 16). Hands-up recall did not change significantly when the band was restricted to 4–20 Hz (−0.85 points; 95% CI −6.21 to +4.51) or 4–13 Hz (−0.45; −7.12 to +6.22), when frontal and temporal channels were excluded (−1.31; −7.51 to +4.88), or with both restrictions (−4.03; −10.49 to +2.44). Overall accuracy fell by 3.39 points when frontal and temporal channels were excluded (95% CI −5.77 to −1.00; Holm-corrected p = 0.019 by t-test and 0.055 by Wilcoxon test) and by 4.33 points with both restrictions (Holm-corrected p = 0.001 and 0.002), with a larger estimated reduction in eyes-closed recall (−6.66 points,

−11.22 to −2.11, and −4.92, −9.66 to −0.17), while hands-up recall estimates remained imprecise. These intervals are wide (about ±6 points) and come from a single partition, so they leave appreciable changes in hands-up recall compatible with the data; they do not establish a cortical origin of the hands-up signal (Section 4).

Condition	Δ accuracy (pp)	Δ hands-up recall (pp)	Δ eyes-closed recall (pp)	Hands-up → eyes-open (%)
Reference (4–40 Hz, 19 channels)	63.02% (reference)	52.0% (reference)	88.9% (reference)	40.6
Band 4–20 Hz	+0.07 (−1.58 to +1.72)	−0.85 (−6.21 to +4.51)	+1.25 (−0.83 to +3.33)	41.4
Band 4–13 Hz	−1.15 (−3.17 to +0.87)	−0.45 (−7.12 to +6.22)	+0.51 (−2.95 to +3.97)	40.5
Without frontal and temporal channels	−3.39 (−5.77 to −1.00)	−1.31 (−7.51 to +4.88)	−6.66 (−11.22 to −2.11)	41.2
Without frontal and temporal channels, 4–20 Hz	−4.33 (−6.59 to −2.06)	−4.03 (−10.49 to +2.44)	−4.92 (−9.66 to −0.17)	43.8

Table 16 *Hands-up sensitivity analysis of the EEG-only EEG-Conformer (one participant partition and seed). Differences from the reference configuration in participant-level accuracy and class recall with 95% confidence intervals; the last column is the pooled proportion of hands-up epochs classified as eyes-open*

4. Discussion

The contribution of this study is methodological and empirical rather than architectural. Its central finding concerns the stability of small effects. In a single participant-wise split, removing the accelerometer channels appeared to improve accuracy by 2.9 points, and on PhysioNet removing the Transformer appeared to reduce it by 2.1 points. When the same comparisons were evaluated with five-fold participant-wise cross-validation repeated five times, with a different partition and initialization seed in each repeat, the accelerometer effect reversed (including the channels helped by 1.2 points) and the PhysioNet Transformer effect disappeared (−0.2 points), while on the custom dataset the Transformer showed a small, fairly consistent benefit (0.7 points). The variation of a participant's accuracy across partitions and seeds (SD about 4 points) is as large as the single-split effects, and the variation between participants (SD 8–9 points) is larger still. A single split can therefore misstate not only the size but also the sign of a component effect, and the dataset-dependent pattern we inferred from three single-split ablations was largely a product of this instability. The same applies to the explanation we first offered for the accelerometer result (a poorly normalized or unexploited signal): because the effect did not replicate, that hypothesis, which we raised after seeing the single-split result, is withdrawn. The accelerometer channels do carry task-related information (participant-level $F(2,94) = 6.33$, $p = 0.0026$), consistent with the small benefit of including them in the repeated analysis, and alternative fusion strategies, such as separate normalization branches, remain untested.

A second finding concerns data leakage and the unit of analysis. On BCI Competition IV-2a, EEG-Conformer accuracy fell from 59.40% to 46.01% (−13.4 points) when the split changed from trial-level to participant-wise, and under full leave-one-subject-out cross-validation no significant difference was found between EEG-Conformer and EEGNet (44.02% versus 44.14%; $p = 0.911$; over five seeds, 44.24% versus 44.40%; $p = 0.810$). Because the trial-level and participant-wise evaluations also differ in their test participants and training-set size, the 13.4-point difference cannot be attributed entirely to leakage, although leave-one-subject-out cross-validation over all nine participants gave a similarly low estimate. On the custom dataset, the EEG-Conformer versus EEGNet advantage (+3.05 to +3.15 points in the original five-fold analysis) was not significant across the five folds ($p = 0.195$) but was significant across the 48 test participants (paired t-test $p = 0.030$; Wilcoxon $p = 0.014$). These tests answer different questions, variability

across training partitions versus generalization across participants, and the participant-level test does not account for participants sharing a trained model within a fold, so its p value is likely optimistic. The same nesting limits the ablation analysis, which reduces the dependence on a single partition but not this dependence, so its p values should also be read as approximate. On PhysioNet, EEG-Conformer's nominal advantage over EEGNet (+1.04 points) was not distinguishable from zero (95% CI −1.12 to +3.20). The direct comparisons between EEG-Conformer and the CNN baselines (Tables 4, 9 and 12) reflect the whole architecture, including its convolutional front end, and should not be attributed to self-attention without the corresponding ablation.

A third finding concerns the acquisition rate. Exploratory analyses detected no clear accuracy reduction after downsampling from 500 to 125 Hz or restricting the band to 4–13 Hz, and keeping only 13–40 Hz cost 5.6 points, mainly through eyes-closed recall (−15.3 points), consistent with a substantial contribution of activity below 13 Hz, including the alpha increase with eye closure, rather than of fine temporal detail. These analyses do not support a decoding advantage from 500 Hz acquisition under the tested conditions; they do not establish equivalence or identify all temporal features used by the model. The Transformer sees about 129 pooled tokens rather than 2000 samples (Section 2.1), so it is not a test of long-sequence modeling at 500 Hz either. Window length mattered modestly (1 s windows lost 2.3 points). Scaling the kernels to comparable physical durations produced no statistically detectable accuracy difference in this exploratory evaluation, but the different kernel durations across sampling rates (Table 3) remain a caveat for direct comparisons between the three datasets.

The per-class results on the custom dataset further clarify what the model is (and is not) learning: eyes-closed decoding is highly reliable, consistent with the strong and well-characterized occipital alpha response, whereas distinguishing hands-up from eyes-open is substantially harder, since both conditions share an open-eyes visual state; whether a cortical sensorimotor contribution is present cannot be separated from muscle, movement and recording-block effects in this design. This suggests that future work on this paradigm may benefit from task designs or feature representations that more directly emphasize sensorimotor rhythms (e.g., mu-band power over C3/C4/Cz) rather than relying on the network to discover this distinction unsupervised from raw, filtered voltage traces. An important caveat applies to the hands-up condition. The task required overt, sustained arm elevation, which produces electromyographic (EMG) activity in neck, shoulder and scalp muscles and postural movement that propagates into scalp EEG, particularly at frontal and temporal electrodes. Several features of the present pipeline make such contamination likely to remain in the data: EMG is broadband and overlaps the upper part of the 4–40 Hz analysis band (beta and low gamma); the dry electrodes used are more susceptible to motion artifacts than gel-based electrodes; the ±500 μV rejection threshold removes only gross artifacts, not low-amplitude muscle activity; and no EMG was recorded and no ICA- or regression-based artifact correction was applied. The accelerometer analysis (Section 3.3) is consistent with greater movement during hands-up than during either resting condition. Consequently, the decodable signal for the hands-up class cannot be attributed to sensorimotor cortical activity, and the three-class accuracies should not be interpreted as evidence of motor decoding from cortical signals. Our sensitivity analyses (Section 3.7) did not show that hands-up recall depended on frequencies above 20 Hz or on frontal and temporal channels, but their intervals are wide (about ±6 points), and a muscle contribution at low frequencies or at central electrodes cannot be excluded. Disambiguating these sources would require concurrent EMG, artifact-correction methods targeting muscle components, or motor-imagery rather than motor-execution paradigms.

The baseline architectures compared in this study (EEGNet, ShallowConvNet, DeepConvNet) and EEG-Conformer itself represent the CNN and CNN-Transformer design families most widely used as reference points in the EEG-decoding literature, but a substantial and rapidly growing body of recent work, including several 2025 studies, has proposed further hybrid CNN-Transformer variants for EEG motor-imagery decoding, such as CNN-based multi-branch fusion architectures (Mallat et al. 2025), CNN-GRU hybrids (Bouchane et al. 2025), and CNN-Swin-Transformer models using hierarchical window-based self-attention (Deng et al. 2025); a recent review (Vafaei and Hosseini 2025) surveys this broader landscape of

transformer architectures (time-series, vision, and graph-attention transformers and their hybrids) applied to EEG. We did not implement these newer models because the purpose of this study was not to establish a new state-of-the-art result, but to evaluate, via ablation and participant-wise cross-validation, whether the architectural components already present in a representative CNN-Transformer hybrid (self-attention and multimodal sensor fusion) provide a measurable benefit; the evaluation procedures used here (participant-wise cross-validation, paired comparisons, and modality/component ablation) are architecture-agnostic and could equally be applied to these newer models in future work. Because no recent baseline was evaluated experimentally, the results do not establish competitiveness with current architectures. Fatigue and affective decoding and time-blocked evaluation were not tested; generalization to these settings requires independent evidence.

This study has several limitations. First, the sample size (48 participants) is modest for deep learning. Both the single 38/10 split (per-participant accuracy 44.59% to 91.89%; 95% CI [57.17%, 77.37%]; Section 3.1) and the 5-fold participant-wise cross-validation (EEG-Conformer 64.68% +/- 3.08%, EEGNet 61.63% +/- 4.05% across folds) showed substantial variability. The EEG-Conformer vs. EEGNet difference was significant only at the participant level, and that test treats participants who share a trained model as independent; each model was also trained once per fold with fixed random seeds, so variability due to random initialization was not estimated for this comparison. The repeated cross-validation described in Section 2.9 was applied to the ablations and to the PhysioNet comparison but not to the model comparisons on the custom dataset, which therefore remain single-partition estimates. A larger sample, or full leave-one-participant-out cross-validation repeated over seeds, would be needed to estimate these differences precisely. Second, the hands-up condition was a motor-execution task without EMG recording or muscle-artifact correction, so its classification may partly rely on EMG and motion artifacts rather than cortical activity (Section 4). Third, although both public-benchmark comparisons (Sections 3.5-3.6) use participant-wise splits, the number of held-out participants in the single-split estimates is small in both cases (10 of 49 for PhysioNet; 2 of 9 for BCI Competition IV-2a); for BCI Competition IV-2a we additionally performed full leave-one-subject-out cross-validation (Section 3.6), which substantially changed the conclusion (no significant EEG-Conformer vs. EEGNet difference, versus EEGNet appearing ahead in the single 7/2 split), underscoring the value of full cross-validation over a single split. The PhysioNet comparison was extended to repeated five-fold participant-wise cross-validation for EEG-Conformer, its CNN-only variant and EEGNet (Section 3.5), but not for ShallowConvNet and DeepConvNet. Fourth, the amplitude-based artifact-rejection threshold (±500 μV, tolerating up to two noisy channels per epoch) used for the custom dataset was chosen based on inspection of that dataset and may require adjustment for other recording setups. Finally, two participants of the custom dataset were excluded because their truncated recordings could not be read by the conversion step (Section 2.6); the content-hash check in that step was computed on the parsed output and therefore flagged unreadable files as duplicates, which hashing the raw files avoids. Other analyses may benefit from validating the same pipeline on data from a second recording site. Because the custom recordings were made in a fixed order (eyes-closed, eyes-open, hands-up) and each state was a separate block, time-dependent factors such as changes in electrode contact or drowsiness are confounded with state and could provide a participant-general shortcut; the same applies to the PhysioNet task, whose three states also come from separate runs. To estimate the time-related signal that is available, we classified the first versus the second half of each recording in held-out participants (balanced accuracy 54.0% [95% CI 51.6–56.5] for eyes-open, 52.8% [50.1–55.5] for eyes-closed and 52.9% [50.4–55.4] for hands-up; chance 50%). These values are only slightly above chance, indicating a weak participant-general time-related signal within recordings (the eyes-closed result was significant under the t-test but not under the Wilcoxon test), but they do not quantify differences between blocks, which cannot be separated from state in this design. The sensitivity analyses (sampling rate, band limit, window length, hands-up) used a single participant partition and seed, so their intervals do not include partition variability. The study is also limited in scope. The leakage and split-dependence problems examined here are not specific to the eyes-open, eyes-closed, hands-up and motor-imagery tasks studied; they are likely more severe in fatigue and affective decoding, where labels are constant over whole sessions or trials and time-blocked validation is required.

We did not test such settings or newer architectures beyond the CNN baselines, so our conclusions are limited to the datasets and models evaluated.

5. Conclusions

Rather than proposing a new architecture, we evaluated the published EEG-Conformer (Song et al. 2023) on a custom 500 Hz dataset from 48 participants (eyes-open, eyes-closed, hands-up) and, with participant-wise splits, on PhysioNet (three states from separate runs) and BCI Competition IV-2a (four-class motor imagery). On the custom dataset the model reached 67.27% on a single split, comparable to EEGNet; across five repeats with different participant partitions and initialization seeds, mean accuracy was 62–64% depending on the configuration. The accelerometer channels and the Transformer had small effects (including them raised accuracy by 1.2 and 0.7 points; Holm-corrected $p = 0.060$ and 0.027), and single-split estimates of these effects (−2.9 points for the accelerometer channels and up to +3.4 points for the Transformer in the EEG-only model) did not predict them. On PhysioNet the single-split Transformer benefit of 2.1 points disappeared in repeated cross-validation (−0.2 points). On BCI Competition IV-2a, leave-one-subject-out cross-validation over five seeds likewise showed no clear Transformer contribution (+0.5 points; 95% CI −1.5 to +2.5) and no clear difference from EEGNet (−0.2 points). Exploratory comparisons detected no clear accuracy reduction after downsampling to 125 Hz or restricting the band to 4–13 Hz, whereas removing frequencies below 13 Hz reduced accuracy. These single-partition results do not establish equivalence and do not support a high-temporal-resolution advantage. Eyes-closed epochs were decoded reliably, consistent with the alpha increase with eye closure, whereas hands-up epochs were classified as eyes-open about 40% of the time, and their origin (cortical versus muscle and movement) could not be resolved. The states were recorded in a fixed order, which cannot be separated from state in this design. Trial-level splitting yielded 13.4 points higher BCI IV-2a accuracy than a participant-wise split, consistent with leakage between trials of the same participant. These results argue for participant-wise evaluation repeated over partitions and seeds, and for treating single-split ablation differences of a few points as unreliable.

Acknowledgements

This work was supported by the Graduate School of Konyang University.

Statements and Declarations

Funding: This work was supported by the Graduate School of Konyang University.

Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics approval: The study was approved by the Institutional Review Board (IRB) of Konyang University (KYU 2022-12-004-001), and was conducted in accordance with the Declaration of Helsinki. The approved protocol, “A Preliminary Study on the Utilization of User EEG Data for Interface Technology Research” (Principal Investigator: Woong Sik Kim, Department of Artificial Intelligence, Konyang University), had an approval period of 25 January 2023 to 24 January 2024, within which the data collection reported in this manuscript was conducted.

Consent to participate: Informed consent was obtained from all individual participants included in the study.

Data availability: The data presented in this study are not publicly available due to privacy restrictions stipulated by the Institutional Review Board of Konyang University (KYU 2022-12-004-001); de-identified data may be made available from the corresponding author upon reasonable request and with permission of the Institutional Review Board of Konyang University.

Author contributions: Conceptualization: Woong Sik Kim; Methodology: Ha Rang Cho, Woong Sik Kim; Software: Ha Rang Cho; Formal analysis: Ha Rang Cho; Data curation: Ha Rang Cho; Visualization: Ha Rang Cho; Resources: Woong Sik Kim; Writing – original draft preparation: Ha Rang Cho; Writing – review and editing: Woong Sik Kim; Supervision: Woong Sik Kim; Project administration: Woong Sik Kim; Funding acquisition: Woong Sik Kim. All authors read and approved the final manuscript.

Use of generative artificial intelligence: described in Section 2.11.

References

Abiri R, Borhani S, Sellers EW, Jiang Y, Zhao X (2019) A comprehensive review of EEG-based brain–computer interface paradigms. J Neural Eng 16:011001

Berger H (1929) Über das Elektrenkephalogramm des Menschen. Arch Psychiatr Nervenkr 87:527–570

Bouchane M, Guo W, Yang S (2025) Hybrid CNN-GRU models for improved EEG motor imagery classification. Sensors 25:1399

Brunner C, Leeb R, Müller-Putz G, Schlögl A, Pfurtscheller G (2008) BCI Competition 2008 – Graz data set A. Graz University of Technology, Graz

Craik A, He Y, Contreras-Vidal JL (2019) Deep learning for electroencephalogram (EEG) classification tasks: a review. J Neural Eng 16:031001

Deng X, Huo H, Ai L, Xu D, Li C (2025) A novel 3D approach with a CNN and Swin Transformer for decoding EEG-based motor imagery classification. Sensors 25:2922

Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101:e215–e220

Kim Y, Han J, Kim W (2024) Effect of data continuity in EEG signal-based BCI research. J Korea Inst Converg Signal Process 25(1) (in Korean)

Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ (2018) EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. J Neural Eng 15:056013

Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, Yger F (2018) A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. J Neural Eng 15:031005

Mallat S, Hkiri E, Albarrak AM, Louhichi B (2025) A synergy of convolutional neural networks for sensor-based EEG brain–computer interfaces to enhance motor imagery classification. Sensors 25:443

Roy Y, Banville H, Albuquerque I, Gramfort A, Falk TH, Faubert J (2019) Deep learning-based electroencephalography analysis: a systematic review. J Neural Eng 16:051001

Schirrmeister RT, Springenberg JT, Fiederer LDJ, Glaser M, Eggensperger K, Tangermann M et al (2017) Deep learning with convolutional neural networks for EEG decoding and visualization. Hum Brain Mapp 38:5391–5420

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, pp 618–626

Song Y, Zheng Q, Liu B, Gao X (2023) EEG Conformer: convolutional Transformer for EEG decoding and visualization. IEEE Trans Neural Syst Rehabil Eng 31:710–719

Vafaei E, Hosseini M (2025) Transformers in EEG analysis: a review of architectures and applications in motor imagery, seizure, and emotion classification. Sensors 25:1293

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al (2017) Attention is all you need. Adv Neural Inf Process Syst 30

Wolpaw JR, Wolpaw EW (eds) (2012) Brain–computer interfaces: principles and practice. Oxford University Press, Oxford