Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 239-262
DOI: https://doi.org/10.51903/jtie.v5i1.493

Calibration-Light Subject-Independent Motor Imagery BCI via Self-Supervised Pretraining and Conformer

Qiyou Wu*[1], Gaotian Mi[2], Dan Wood[3]
Email: qiyou.wu0106@outlook.com
[1]*Artificial Intelligence, Northeastern University, MA, USA*
[2]*Biomedical Engineering, Johns Hopkins University, MD, USA*
[3]*Computer Engineering, Dartmouth College, NH, USA*
*Corresponding Author

Abstract

Motor imagery (MI) electroencephalography (EEG) is a foundational paradigm for non-invasive brain–computer interfaces (BCIs). However, its practical adoption is constrained by time-consuming per-user calibration and limited cross-subject generalization. This study evaluates a calibration-light MI-BCI framework that combines self-supervised masked EEG pretraining with a lightweight Conformer fine-tuning model. Experiments were conducted on BCI Competition IV Dataset 2b using only the labeled sessions 01T–03T, with artifact-annotated trials removed according to the official 1023 markers. Three deployment-relevant settings were examined: within-subject evaluation (01T–02T → 03T), strict leave-one-subject-out (LOSO) evaluation, and few-shot adaptation with k = 1/5/10 trials per class from the held-out subject's screening sessions. Full within-subject benchmarking included CSP+LDA, EEGNet, DeepConvNet, ShallowFBCSPNet, supervised Conformer, and SSL+Conformer, while the subject-independent and few-shot analyses focused on CSP+LDA, EEGNet, supervised Conformer, and SSL+Conformer. In the fully calibrated setting, the best mean accuracy was obtained by ShallowFBCSPNet (62.23% ± 14.16%), whereas SSL+Conformer achieved 54.85% ± 11.15% and slightly outperformed the supervised Conformer (53.56% ± 8.81%). Under strict LOSO, EEGNet achieved the highest mean accuracy (52.92% ± 8.25%), while SSL+Conformer reached 51.56% ± 7.18%. In few-shot adaptation, SSL+Conformer achieved the highest mean accuracy at k = 10 (52.84% ± 7.64%) among the core calibration-light methods. The proposed model had a size of 0.1329 MB, a median CPU latency of 0.8777 ms/trial, and LOSO calibration values of ECE = 0.0630 and Brier = 0.4995. These results indicate that masked EEG pretraining provides a competitive lightweight baseline and is most useful when a modest amount of target-subject calibration data is available.

Keywords: *Motor Imagery, EEG, Brain–Computer Interface, Subject-Independent Learning.*

I. INTRODUCTION

Motor imagery (MI) BCIs decode the mental rehearsal of movements from EEG to enable communication and control without overt muscular activity (Handoko et al., 2025; Nita et al., 2025; Roni et al., 2025). Since the early demonstrations of sensorimotor rhythm modulation for BCI control, MI has become one of the most widely studied paradigms due to its non-invasive nature, relatively simple experimental procedures, and compatibility with low-cost hardware (Pfurtscheller & Neuper, 2001). In standard MI tasks, subjects imagine left- or right-hand movements, producing class-dependent changes in mu (8–13 Hz) and beta (13–30 Hz) rhythms over the sensorimotor cortex. These oscillatory changes can be captured from a small number of EEG channels (e.g., C3/C4/Cz) and are amenable to time–frequency and spatial filtering approaches.

Received on Januari 2025; Revised on February 2025; Accepted on April 2025; Published on April 2025

Despite decades of methodological development, the practical deployment of MI-BCIs remains difficult. A central barrier is calibration: most systems require a dedicated, subject-specific data-collection session to tune spatial filters and classifiers. Calibration time is a major usability bottleneck for real-world BCI adoption because new users must repeatedly perform MI trials before the system becomes functional. From a human–computer interaction perspective, reducing calibration reduces cognitive fatigue and improves accessibility. From a clinical perspective, calibration can be particularly burdensome for users with limited attention span or motor impairment.

A second barrier is cross-subject generalization. MI EEG exhibits large inter-individual variability in rhythmic peak frequencies, spatial topographies, and nonstationary session-to-session drift. Even when a model performs well within-subject, it often degrades when applied to a new subject without calibration. This shift resembles a domain adaptation problem in which each subject (or session) constitutes a separate domain. The BCI Competition series was established to benchmark algorithms under standardized conditions and to promote reproducible comparisons. In BCI Competition IV, multiple MI datasets were released and have since become canonical benchmarks for MI decoding (Tangermann et al., 2012).

BCI Competition IV Dataset 2b (Graz dataset B) is particularly useful for studying calibration-light learning because it includes nine subjects, two MI classes, and multiple sessions that separate screening (without feedback) and feedback (online) recordings (Leeb et al., 2008). The dataset is small enough to enable extensive algorithmic comparison yet large enough to expose generalization issues. In this study, we restrict the analysis to the labeled sessions 01T–03T and exclude the evaluation sessions 04E/05E. This restriction aligns with reproducible open research uses and matches the explicit study requirement.

In practice, “calibration-light” has multiple operational meanings. In some settings, calibration-light refers to a purely subject-independent model that uses no labeled data from the new user; in other settings, it refers to few-shot calibration, where a small number of labeled trials are collected and used for rapid adaptation. These definitions correspond to distinct deployment modes: immediate out-of-the-box operation versus minimal interactive calibration. This manuscript therefore evaluates both strict subject-independent generalization and few-shot adaptation, explicitly separating these settings to avoid ambiguous claims.

MI decoding is typically embedded in a control loop that imposes engineering constraints. Many BCI applications provide real-time feedback to the user, and a delay in the loop can impair learning and control stability. The control loop also depends on probabilistic outputs: thresholds, rejection rules, and adaptive interfaces use class probabilities rather than hard labels. For these

reasons, this study includes both inference latency and calibration metrics as first-class evaluation outcomes. A model that is accurate but overconfident can produce unstable control, and a model that is accurate but slow can fail real-time constraints.

A further practical constraint is compute budget. Many research prototypes use desktop GPUs, but deployed BCIs can run on laptops, tablets, or embedded devices. Consequently, approaches that require large models or expensive per-user optimization may not transfer to real systems. The present work therefore emphasizes a lightweight Conformer configuration. It includes model size and CPU latency measurements as explicit deliverables, enabling readers to assess the feasibility of real-time deployment alongside decoding performance.

CSP-based spatial filters and linear classifiers dominate classical MI decoding. CSP optimizes spatial projections that maximize variance differences between classes, and when combined with log-variance features and LDA, it yields strong performance in many within-subject settings (Blankertz et al., 2008). However, CSP is highly dependent on training covariance estimates and therefore sensitive to cross-subject and cross-session shifts. Deep learning has reduced reliance on handcrafted features by learning temporal and spatial filters directly from raw EEG. EEGNet, DeepConvNet, and ShallowFBCSPNet are widely used architectures that perform competitively across BCI paradigms (Lawhern et al., 2018; Schirrmeister et al., 2017). Nevertheless, deep networks trained from scratch on small per-subject datasets can overfit and often require careful regularization, especially when transferring across subjects.

Self-supervised learning (SSL) offers an alternative to purely supervised training by leveraging unlabeled EEG to learn transferable representations. Masked modeling, in which parts of the input are removed, and the model is trained to reconstruct them, is one of the most successful SSL strategies in modern representation learning (He et al., 2022). For EEG, masked modeling can encourage an encoder to capture temporal dependencies and cross-channel structure without requiring labels. These representations can then be adapted with limited labeled MI trials, supporting calibration-light usage.

Transformers provide a flexible framework for modeling long-range dependencies, but can be data-hungry and may underperform when training data are limited. Conformer architectures integrate convolutional and self-attention operations, combining locality priors with global context modeling. Conformer was originally proposed for speech recognition (Gulati et al., 2020) and is increasingly adopted in biosignal modeling, as biosignals exhibit both short-term structure and long-range temporal dependencies. In MI EEG, convolutional modules help capture oscillatory envelopes and frequency-localized patterns, while attention modules can capture temporal context and global dependencies within an imagery window.

This study investigates whether masked EEG modeling pretraining can improve Conformer-based MI decoding under leakage-free calibration-light evaluation. The core contributions are as follows. (1) It implements a lightweight attention–convolution encoder for low-channel MI EEG and evaluates it on BCI Competition IV Dataset 2b. (2) It compares within-subject, strict LOSO, and few-shot calibration protocols with explicit session usage and leakage controls. (3) It jointly reports discrimination, calibration, and engineering metrics. (4) It provides per-subject analyses, paired statistical comparisons, and ablations on mask ratio and conformer depth.

The remainder of the manuscript is organized into a focused literature review, methods, results, discussion, and conclusion. Because reproducible benchmarking is a primary objective in MI-BCI research, emphasis is placed on consistent dataset usage, leakage-free protocol design, and complete reporting.

II. LITERATURE REVIEW

MI decoding has a long tradition of signal processing methods that exploit rhythmic modulation. A canonical pipeline applies band-pass filtering, extracts band-power or time–frequency features, and learns spatial filters to maximize discriminability between MI classes. CSP is among the most influential spatial filtering techniques, producing projections that maximize the ratio of class-specific variances under covariance constraints. CSP-based features are often fed into LDA due to its robustness and efficiency (Blankertz et al., 2008). Variants such as filter-bank CSP (FBCSP) apply CSP over multiple frequency bands, then select informative band–component features to improve robustness. A broad review of BCI classification algorithms highlights that linear classifiers coupled with physiologically motivated features can be highly effective, but that performance depends strongly on subject-specific calibration (Lotte et al., 2007).

CSP can be interpreted as a supervised spatial whitening transform that emphasizes variance patterns that differ across classes. Because CSP is derived from covariance matrices estimated on the training set, it is sensitive to sample size and to nonstationary noise. Regularization of covariance estimates, shrinkage, and robust covariance estimators can partially mitigate instability. However, even with regularization, CSP filters learned on one subject often transfer poorly to another subject because the underlying sensorimotor rhythms and their topographical projections differ across individuals. This property motivates subject-invariant representation learning approaches that move beyond explicit covariance fitting per subject.

A key limitation of classic pipelines is their reliance on hand-selected frequency bands. While mu and beta rhythms are central for MI, peak frequencies vary across subjects and can drift across sessions. Consequently, fixed filter bands can be suboptimal. Filter-bank approaches address this by including multiple overlapping bands, but they expand the feature space and can increase

overfitting when labels are limited. Deep learning addresses band selection by learning temporal filters, but deep models can overfit in small-data regimes and be sensitive to preprocessing choices.

Deep learning has substantially influenced EEG decoding by learning features end-to-end from minimally processed signals. Schirrmeister et al. (2017) demonstrated that both deep and shallow convolutional networks can decode EEG across tasks, and that interpretability analyses reveal learned filters that resemble spectral band-power features. In particular, ShallowFBCSPNet was designed to mimic FBCSP-like operations (temporal filtering, spatial filtering, nonlinear transforms, and pooling), providing a deep learning analogue of classic MI pipelines. Deep ConvNet provides greater capacity and can capture more complex patterns, but may require larger datasets and stronger regularization.

The interpretability of deep EEG models remains an important research topic because BCIs benefit from physiological plausibility. In MI tasks, successful models often attend to patterns consistent with event-related desynchronization/synchronization in sensorimotor regions. Interpretability tools such as activation maximization, saliency maps, and perturbation analysis can identify whether models rely on physiologically meaningful features or on spurious correlations. While interpretability is not the primary focus of this manuscript, the architecture selection (ShallowFBCSPNet and Conformer with a conv stem) reflects a preference for inductive biases that align with established MI signal structure.

EEGNet (Lawhern et al., 2018) introduced a compact convolutional architecture that uses depthwise and separable convolutions to reduce parameters and improve generalization on small datasets. EEGNet has become a standard baseline because it performs competitively across BCI paradigms while being efficient enough for real-time use. Compared with larger CNNs, EEGNet's inductive bias toward learning temporal filters followed by spatial filters aligns with neurophysiological expectations: temporal filtering captures frequency-selective rhythms, and spatial filtering captures sensorimotor topography. Nevertheless, EEGNet and other CNNs trained from scratch can still struggle in subject-independent settings, where the data distribution of a new subject differs from training subjects.

Cross-subject transfer in MI-BCI is commonly framed as a transfer-learning or domain-adaptation problem because training and test subjects follow related but shifted distributions. Reviews of short/zero-calibration BCIs and MI transfer learning show that alignment-based, reweighting, and subject-invariant representation-learning strategies can reduce the calibration burden (Ko et al., 2021; Li & Xu, 2024). Representative MI studies include weighted transfer learning, Euclidean alignment, and Wasserstein-based domain adaptation (Azab et al., 2019; He & Wu, 2020; She et

al., 2023). In EEG, domain shifts arise from physiological variability, differences in electrode placement, and session nonstationarity. Few-shot calibration remains especially relevant because a small amount of labeled target data may be feasible in practice, whereas full calibration is not.

Domain adaptation methods differ in their assumptions about the availability of data from the target subject. Unsupervised domain adaptation assumes unlabeled target data, while supervised or semi-supervised adaptation assumes some labeled target data. In BCIs, unlabeled target data may be available if the user is willing to perform trials without ground-truth labels. However, fully unlabeled adaptation is limited by the inability to evaluate correctness. Few-shot adaptation provides a practical compromise: it uses a small number of labeled trials to anchor adaptation while keeping calibration burdens low. Because the data budget is limited, few-shot adaptation benefits from a strong representation learned from other subjects, motivating SSL pretraining.

Transformers have become an important family of models for EEG decoding because self-attention can capture temporal dependencies and cross-channel interactions. However, pure attention models can be data-hungry, especially on small MI datasets. Recent MI-EEG studies therefore favor hybrid designs that combine convolution with attention to balance local inductive bias and global context modeling (Wimpff et al., 2024; Juan et al., 2024). Conformer, originally proposed for speech recognition (Gulati et al., 2020), follows this logic by integrating convolutional modules and self-attention, making it a reasonable architecture for low-channel MI EEG.

Self-supervised learning provides a pathway to reduce labeled-data requirements and improve transferability. While masked autoencoders were first popularized in vision (He et al., 2022), recent EEG work has adapted masked reconstruction to learn transferable biosignal representations and to support cross-individual pretraining (Cai & Zeng, 2024; Fu et al., 2024). For MI EEG, masked modeling is attractive because it avoids heavy handcrafted augmentations and can be paired naturally with few-shot fine-tuning. Unlike contrastive learning, masked modeling does not require carefully designed augmentations, which can be challenging for EEG because augmentations may distort physiologically meaningful information.

A practical advantage of reconstruction-based SSL in EEG is that it directly enforces information preservation. When epochs are masked, the encoder must infer missing content from surrounding context, which can encourage the model to represent oscillatory phase and amplitude envelopes as well as cross-channel relationships. However, reconstruction losses also impose an inductive bias toward signal-level fidelity, which may capture noise. Therefore, masked modeling must be paired with downstream evaluation and ablation analyses, such as varying mask ratios and patch sizes, to validate that the learned representations improve MI discrimination and calibration.

In addition to discrimination performance, calibration of probabilistic outputs is important. Guo et al. (2017) highlighted that modern neural networks can be miscalibrated—often overconfident—even when accurate. In BCIs, miscalibration can cause unstable control and can degrade adaptive feedback systems. Expected Calibration Error (ECE) summarizes the mismatch between predicted probabilities and empirical accuracy across bins, and the Brier score provides a proper scoring rule for probability forecasts (Brier, 1950). Reliability diagrams visualize calibration and can complement ECE by revealing systematic over- or under-confidence. Including calibration metrics is therefore aligned with deployment settings where uncertainty estimation and rejection options are important.

Calibration is also connected to user feedback. In online BCIs, feedback is often derived from classifier confidence or smoothed probabilities. Overconfident models can produce misleading feedback, slowing user learning and degrading long-term performance. Conversely, well-calibrated probabilities support stable thresholding, selective classification, and error-aware adaptation. Post-hoc calibration methods such as temperature scaling can improve probability quality without changing the decision boundary, making them attractive for BCI systems that require stable control policies (Guo et al., 2017).

BCI decoding models must also cope with nonstationarity across sessions and users. In practical MI-BCI systems, recalibration, confidence thresholding, and selective rejection are commonly used to maintain stable control when signal distributions drift. This connection again highlights why probability calibration matters: when predicted probabilities more accurately reflect empirical accuracy, downstream control policies can be designed more conservatively and transparently.

BCI performance reporting must also account for chance levels and statistical significance. In small-sample MI studies, raw accuracy can fluctuate substantially, so chance-corrected metrics and paired subject-level tests are important. Müller-Putz et al. (2008) emphasized that “better-than-random” claims should be interpreted in light of trial counts and uncertainty. Accordingly, this study reports Cohen’s kappa, along with accuracy, and uses paired Wilcoxon signed-rank tests across subjects.

Finally, reproducible evaluation requires standardized benchmarks and transparent protocols. The BCI Competition IV review emphasized consistent data splits and careful interpretation across datasets (Tangermann et al., 2012). Recent MI-EEG surveys similarly note that preprocessing, split definition, and calibration protocol can dominate reported performance, which makes leakage-free benchmarking essential (Ko et al., 2021; Wang et al., 2024). Accordingly, the present

study explicitly restricts the analysis to the labeled sessions 01T–03T, excludes 04E/05E, and reports the exact settings for within-subject, LOSO, and few-shot evaluations.

III. RESEARCH METHOD

This section describes the overall methodology used in this study, including the conceptual pipeline, dataset characteristics, preprocessing procedures, evaluation settings, and the decoding methods compared. The proposed calibration-light MI-BCI framework is first illustrated through the conceptual processing pipeline and model architecture.

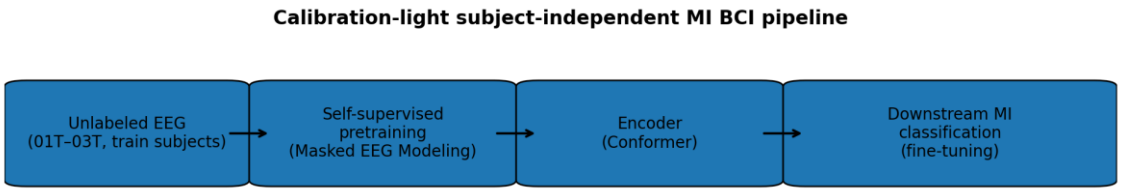

Figure 1. Calibration-light MI-BCI Pipeline (Conceptual)

Specifically, the overall system workflow is summarized in Figure 1, followed by the Conformer fine-tuning architecture in Figure 2 and the masked-EEG modeling objective for self-supervised representation learning in Figure 3. The evaluation protocols adopted in this study, including within-subject training, cross-subject leave-one-subject-out (LOSO), and few-shot adaptation scenarios, are illustrated in Figure 4.

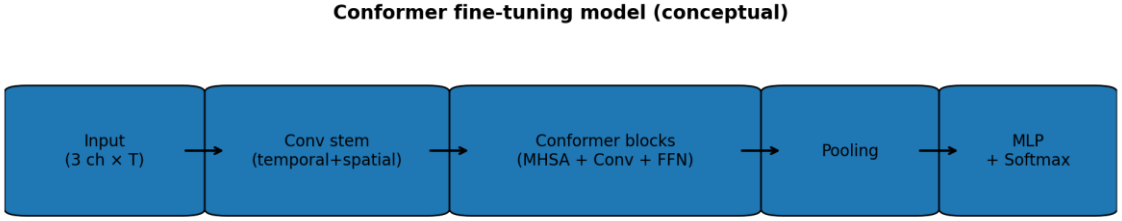

Figure 2. Conformer Fine-Tuning Model (Conceptual)

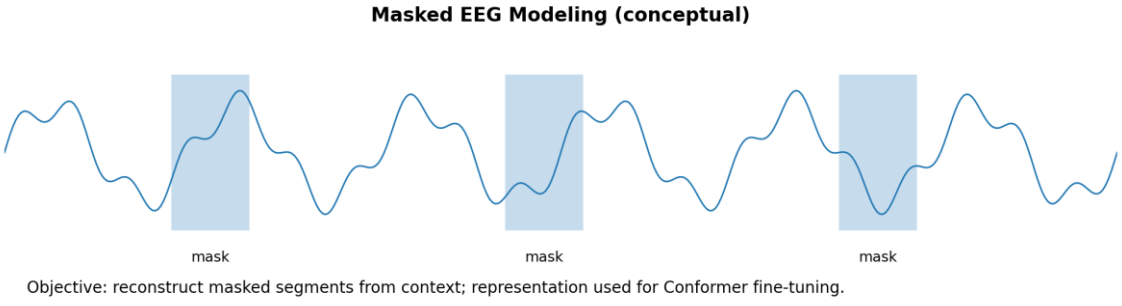

Figure 3. Masked EEG Modeling Objective for Self-Supervised Pretraining (Conceptual)

In addition, the dataset characteristics used in the experiments are summarized in Table 1, while the preprocessing pipeline and signal preparation steps are detailed in Table 2. The experimental

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 239-262
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

evaluation settings and data splits are described in Table 3, and the compared baseline and proposed methods, along with their main components, are summarized in Table 4.

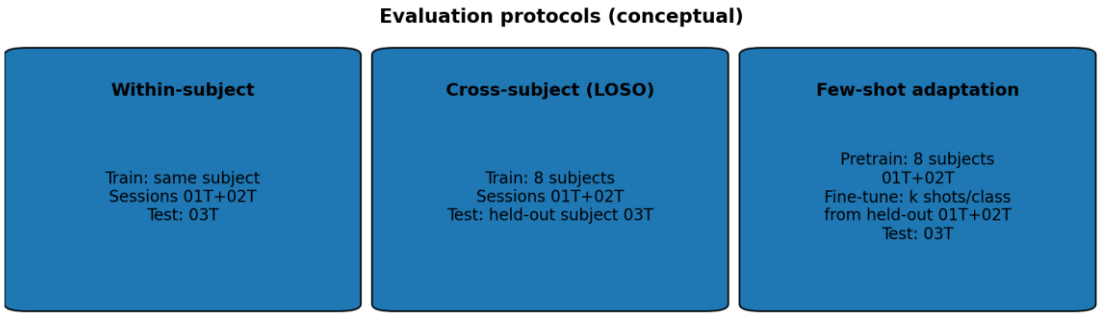

Figure 4. Evaluation Protocols: Within-Subject, Cross-Subject LOSO, and Few-Shot Adaptation (Conceptual)

To keep the calibration-light comparisons focused and computationally consistent, DeepConvNet and ShallowFBCSPNet are reported as additional full-calibration baselines in the within-subject setting, whereas the strict subject-independent LOSO and few-shot analyses focus on CSP+LDA, EEGNet, supervised Conformer, and SSL+Conformer.

Table 1. Dataset Characteristics and Study Usage (BCI Competition IV 2b)

Attribute	Value
Dataset	BCI Competition IV Dataset 2b (Graz data set B)
Subjects	9
Sessions per subject	5 (01T, 02T screening; 03T, 04E, 05E feedback)
Classes	2 (left-hand MI; right-hand MI)
EEG channels	3 bipolar channels: C3, Cz, C4
EOG channels	3 monopolar channels (recorded; excluded from classification)
Sampling frequency	250 Hz
Trials per screening session	120 (6 runs × 20 trials; balanced classes)
Trials per feedback session	160 (4 runs × 40 trials; balanced classes)
Labeled sessions	01T–03T (labels for all trials)
Ignored sessions	04E/05E (ignored in this study)
Cue/imagery timing	Screening: cue at 2 s, imagery from 3–7 s; Feedback: cue from 3–7.5 s

Table 2. Preprocessing Pipeline Specification

Step	Specification
Channel selection	Use EEG channels only (C3, Cz, C4); discard EOG channels
Referencing	Use bipolar channels as provided
Resampling	Downsample from 250 Hz to 125 Hz (factor 2)
Band-pass filtering	4–38 Hz zero-phase Butterworth (4th order)
Epoch window	3.0–7.0 s relative to trial start (4 s imagery period)
Artifact handling	Exclude trials marked with event 1023 in the GDF annotations; then apply per-epoch z-scoring to the retained trials
Normalization	Per-epoch z-score using epoch mean and std over time

Table 5 specifies the key architectural hyperparameters used for each neural model. The table lists the convolutional kernel sizes, the number of filters, the attention configuration, and the classifier

structure used in the experiments. These settings ensure that the compared models maintain comparable capacity while remaining computationally efficient for EEG decoding.

Table 3. Evaluation Protocols and data Splits

Setting	Data split	Goal
Within-subject	Train: subject's 01T+02T; Test: subject's 03T	Subject-specific supervised training
Cross-subject (LOSO)	Train: 8 subjects' 01T+02T; Test: held-out subject's 03T	Strict subject-independent generalization
Few-shot (k=1)	Pretrain: 8 subjects' 01T+02T; Fine-tune: held-out subject k trials/class from 01T+02T; Test: 03T	Minimal calibration
Few-shot (k=5)	Same as above with 5 trials/class	Light calibration
Few-shot (k=10)	Same as above with 10 trials/class	Moderate calibration

In addition, the self-supervised pretraining configuration used in the proposed approach is summarized in Table 6, including the masking strategy, token length, optimization parameters, and training schedule. The evaluation metrics used to assess decoding performance and probability calibration are formally defined in Table 7. These metrics include classification accuracy, Cohen's kappa, macro-F1 score, expected calibration error (ECE), and the Brier score, which together provide complementary perspectives on predictive performance and confidence reliability.

Table 4. Compared Methods and Brief Descriptions

Method	Core components	Rationale
CSP+LDA	Log-variance of CSP-projected signals; LDA classifier	Classic MI baseline; fast; interpretable
EEGNet	Compact CNN with depthwise-separable convolutions	Strong parameter-efficient deep baseline
DeepConvNet	Deep CNN with multiple conv-pool blocks	High-capacity temporal-spatial feature learning
ShallowFBCSPNet	Shallow CNN designed to mimic FBCSP band-power	Well-matched to MI rhythms
Conformer (supervised)	Conv stem + Conformer blocks + classifier	Attention-based deep baseline
SSL + Conformer	Masked EEG modeling pretraining + supervised fine-tuning	Proposed calibration-light approach

Table 5. Neural Network Architecture Hyperparameters (Specified)

Model	Key hyperparameters
EEGNet	Temporal conv: 8 filters, kernel 64; depthwise spatial: D = 2; separable conv: 16 filters; dropout 0.5
DeepConvNet	4 conv blocks (25/50/100/200 filters); first temporal kernel 25; max-pooling kernel 3; dropout 0.5
ShallowFBCSPNet	Temporal conv: 40 filters, kernel 25; spatial conv across channels; squaring + log; avg-pooling kernel 75, stride 15; dropout 0.5
Conformer	Conv stem stride 8; embedding dim 32; 1 Conformer block; 4 attention heads; conv kernel 15; dropout 0.2
Classifier head	Global average pooling + 2-layer MLP (32→32→2)

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 239-262
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

To ensure rigorous comparison across methods, the statistical analysis plan is summarized in Table 8. The analysis reports per-subject performance, aggregated statistics across subjects, and paired statistical tests using the Wilcoxon signed-rank test with Holm correction for multiple comparisons. Effect sizes are also reported to quantify the magnitude of performance differences between methods. Reproducibility considerations are documented in Table 9, including fixed random seeds, deterministic data splits, software versions, hardware reporting, and stored prediction artifacts that enable recomputation of all metrics.

Table 6. Self-Supervised Pretraining Configuration (Masked EEG Modeling)

Component	Setting
Pretext task	Masked EEG modeling (reconstruct masked temporal patches)
Mask ratio	0.5 in the main experiments; ablation at 0.3/0.5/0.7
Token/patch length	8 samples at 125 Hz (≈64 ms)
Loss	Mean squared error on masked patches
Optimizer	AdamW (weight decay 0.01)
Learning rate	1e-3 pretraining; 5e-4 fine-tuning
Batch size	Within-subject: 64 pretraining / 32 fine-tuning; LOSO/few-shot: 256 pretraining / 128 fine-tuning
Epochs	Within-subject: up to 4 pretraining + 8 fine-tuning; LOSO/few-shot: up to 3 pretraining + 4 fine-tuning; early stopping on validation split
Augmentations	None beyond masking

The experimental results are summarized in several tables corresponding to the evaluation settings defined earlier. Within-subject performance across all nine subjects is reported in Table 10, while the strict cross-subject LOSO evaluation is summarized in Table 11. Few-shot adaptation results for different calibration sizes (k = 1, 5, and 10 trials per class) are presented in Table 12. Table 13 reports the key paired statistical comparisons for representative conditions, including corrected p-values and paired effect sizes. In addition to decoding performance, engineering metrics relevant for real-time BCI deployment are summarized in Table 14, including model size and inference latency measured in a controlled CPU environment. Finally, representative ablation results investigating the impact of masking ratio and encoder depth in the proposed SSL framework are summarized in Table 15.

Table 7. Evaluation Metrics and Operational Definitions

Metric	Definition
Accuracy	Proportion of correctly classified trials
Cohen's kappa	Chance-corrected agreement between predictions and labels
Macro-F1	Class-averaged F1 score (robust to imbalance)
ECE	Expected Calibration Error with 15 equal-width bins
Brier score	Mean squared error between predicted probabilities and one-hot labels
Latency	Median CPU forward-pass time per trial (batch size 1)
Model size	Serialized parameter size in megabytes (MB)

The reliability of probabilistic predictions is further examined using reliability diagrams, which visualize the relationship between predicted confidence and empirical accuracy. Reliability diagrams provide insight into whether model confidence estimates are well calibrated,

complementing the numerical calibration metrics reported earlier. In a perfectly calibrated model, predictions with a given confidence level should correspond to the same empirical accuracy. Deviations from the diagonal line indicate overconfidence or underconfidence in the predicted probabilities.

Table 8. Statistical Analysis Plan for Method Comparisons

Item	Specification
Per-subject reporting	Report metrics for each subject and setting
Aggregates	Mean ± standard deviation across subjects
Primary significance test	Wilcoxon signed-rank test on subject-level accuracies
Multiple comparisons	Holm correction within each predefined comparison set
Effect size	Paired Cohen's d
Primary confirmatory comparisons	SSL+Conformer vs CSP+LDA, EEGNet, and supervised Conformer in LOSO and few-shot settings

The reliability diagram for the strict cross-subject evaluation is presented in Figure 5, illustrating calibration behavior when models are tested on unseen subjects without any calibration data. In addition, the reliability diagram for the few-shot adaptation scenario with k = 10 k=10 trials per class is shown in Figure 6, providing a visual comparison of calibration quality after limited subject-specific fine-tuning.

Table 9. Reproducibility Checklist and Reporting Items

Aspect	Details
Random seeds	Fixed seeds for numpy/torch and deterministic dataloading
Split determinism	Documented subject/session splits and few-shot sampling indices
Software	Python 3.11; PyTorch (CPU); numpy; scipy; python-docx
Hardware reporting	CPU model; RAM; OS; single-thread vs multi-thread settings
Training logs	Store loss curves, best epoch, and checkpoint hashes per run
Artifacts	Save per-subject predictions to enable recomputing all metrics and plots

A. Dataset and inclusion criteria

The target dataset is BCI Competition IV Dataset 2b (Graz data set B), a two-class MI dataset recorded at Graz University of Technology. The dataset provides bipolar EEG from channels C3, Cz, and C4, as well as three monopolar EOG channels. Each subject was recorded in five sessions: two screening sessions without feedback (01T and 02T) and three sessions with online feedback (03T, 04E, 05E) (Leeb et al., 2008).

Table 10. Within-Subject Summary (Mean ± SD Across 9 Subjects)

Method	Acc	Kappa	Macro-F1	ECE	Brier
CSP+LDA	58.03 ± 6.70	0.1619 ± 0.1360	0.5701 ± 0.0686	0.0869 ± 0.0315	0.4893 ± 0.0483
EEGNet	52.16 ± 7.33	0.0472 ± 0.1394	0.4895 ± 0.0957	0.0634 ± 0.0354	0.4966 ± 0.0147
DeepConvNet	53.91 ± 11.34	0.0862 ± 0.2207	0.4388 ± 0.1679	0.0771 ± 0.0509	0.4871 ± 0.0645
ShallowFBCSPNet	62.23 ± 14.16	0.2452 ± 0.2818	0.5821 ± 0.1804	0.1342 ± 0.0800	0.4587 ± 0.1724
Conformer (supervised)	53.56 ± 8.81	0.0745 ± 0.1751	0.4742 ± 0.1225	0.0643 ± 0.0537	0.4931 ± 0.0179
SSL + Conformer	54.85 ± 11.15	0.0966 ± 0.2220	0.4958 ± 0.1345	0.0757 ± 0.0911	0.4932 ± 0.0173

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 239-262
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Screening sessions contain 120 trials per session arranged as six runs of 20 trials. Feedback sessions contain 160 trials per session arranged as four runs with 20 trials per class per run. Per dataset guidance, EOG channels are provided to monitor ocular artifacts but must not be used for classification. In this study, only the labeled sessions 01T–03T are used and sessions 04E/05E are excluded.

Table 11. LOSO Summary (Mean ± SD Across 9 Held-Out Subjects)

Method	Acc	Kappa	Macro-F1	ECE	Brier
CSP+LDA	51.76 ± 10.41	0.0379 ± 0.2077	0.4634 ± 0.1233	0.0794 ± 0.0728	0.5016 ± 0.0070
EEGNet	52.92 ± 8.25	0.0592 ± 0.1628	0.5219 ± 0.0795	0.0729 ± 0.0330	0.4979 ± 0.0090
Conformer (supervised)	51.72 ± 6.12	0.0278 ± 0.1206	0.4656 ± 0.0799	0.0417 ± 0.0424	0.4998 ± 0.0056
SSL + Conformer	51.56 ± 7.18	0.0282 ± 0.1442	0.4866 ± 0.0872	0.0630 ± 0.0351	0.4995 ± 0.0051

Table 12. Few-Shot Summary (Mean ± SD Across 9 Held-Out Subjects)

Setting	Method	Acc	Kappa	Macro-F1	ECE	Brier
k = 1	CSP+LDA	51.72 ± 8.15	0.0295 ± 0.1649	0.4764 ± 0.0944	0.0602 ± 0.0568	0.5015 ± 0.0056
k = 1	EEGNet	53.41 ± 7.78	0.0752 ± 0.1524	0.5221 ± 0.0850	0.0991 ± 0.0434	0.5109 ± 0.0435
k = 1	Conformer (supervised)	51.64 ± 7.47	0.0268 ± 0.1502	0.4768 ± 0.1035	0.1373 ± 0.0837	0.5462 ± 0.0812
k = 1	SSL + Conformer	51.55 ± 7.18	0.0290 ± 0.1466	0.4779 ± 0.0891	0.1176 ± 0.0682	0.5263 ± 0.0565
k = 5	CSP+LDA	51.74 ± 7.79	0.0304 ± 0.1572	0.4800 ± 0.0875	0.0601 ± 0.0517	0.5015 ± 0.0057
k = 5	EEGNet	53.17 ± 8.29	0.0592 ± 0.1669	0.5251 ± 0.0854	0.1137 ± 0.0397	0.5105 ± 0.0439
k = 5	Conformer (supervised)	50.74 ± 10.18	0.0164 ± 0.2026	0.5014 ± 0.1017	0.1345 ± 0.0852	0.5364 ± 0.0779
k = 5	SSL + Conformer	50.26 ± 7.57	0.0019 ± 0.1526	0.4878 ± 0.0844	0.1210 ± 0.0880	0.5321 ± 0.0768
k = 10	CSP+LDA	52.40 ± 8.70	0.0436 ± 0.1754	0.4894 ± 0.0990	0.0667 ± 0.0586	0.5013 ± 0.0059
k = 10	EEGNet	49.12 ± 7.13	-0.0216 ± 0.1402	0.4791 ± 0.0688	0.1284 ± 0.0477	0.5254 ± 0.0329
k = 10	Conformer (supervised)	50.21 ± 7.99	0.0130 ± 0.1559	0.4775 ± 0.0947	0.1564 ± 0.0314	0.5366 ± 0.0589
k = 10	SSL + Conformer	52.84 ± 7.64	0.0531 ± 0.1533	0.4881 ± 0.0848	0.1313 ± 0.0662	0.5335 ± 0.0573

B. Trial structure and labeling

In screening sessions, each trial begins with a fixation cross and an auditory beep, followed by a visual cue (left or right arrow) indicating the imagery class. The cue appears after 2 seconds, and the subject performs imagery during the subsequent period. In feedback sessions, the cue is presented for a longer duration and a smiley feedback signal is shown to indicate classifier output. Labels correspond to left-hand versus right-hand imagery. Because the study uses session 03T as the primary test session, it includes both screening-like and feedback-like trial timing in evaluation while maintaining label availability. All label mappings and trial counts are fixed, and any excluded trials are logged to ensure reproducibility.

C. Data representation

After downsampling to 125 Hz and extracting a 4-second imagery window, each trial is represented as a tensor of shape (C=3, T=500). This fixed-length representation supports efficient

batching and consistent convolutional receptive fields across methods. For classical baselines, the same windowed signals are used to compute covariance matrices and CSP features. For deep models, the raw time series is the input, enabling end-to-end learning of temporal and spatial filters.

Table 13. Key Paired Statistical Comparisons (Accuracy)

Setting	Comparison	p	p_Holm	Cohen_d
Within-subject	SSL + Conformer vs CSP+LDA	0.9355	1.0000	-0.2357
LOSO	SSL + Conformer vs CSP+LDA	0.7148	1.0000	-0.0138
Few-shot k = 10	SSL + Conformer vs CSP+LDA	0.6328	0.6328	0.0332
Few-shot k = 10	SSL + Conformer vs EEGNet	0.0547	0.1094	0.5605
Few-shot k = 10	SSL + Conformer vs Conformer (supervised)	0.0195	0.0586	0.8021

Table 14. Engineering Metrics

Method	Model size (MB)	Median latency (ms)	Latency IQR (ms)
ShallowFBCSPNet	0.0353	0.3888	0.033
EEGNet	0.0136	0.4516	0.1369
SSL + Conformer	0.1329	0.8777	0.0653
Conformer (supervised)	0.1329	0.9003	0.1129
DeepConvNet	1.0343	0.9885	0.0813

Table 15. Representative Ablation Results

Ablation	Representative held-outs	Mean acc
Mask ratio 0.3 (depth = 1)	B01, B04, B08	57.07%
Mask ratio 0.5 (depth = 1)	B01, B04, B08	56.97%
Mask ratio 0.7 (depth = 1)	B01, B04, B08	56.29%
Depth = 1 (mask ratio 0.5)	B01, B04, B08	56.97%
Depth = 2 (mask ratio 0.5)	B01, B04, B08	51.30%

D. Preprocessing Data

Preprocessing follows Table 2. In addition to discarding EOG channels, trials marked with event 1023 in the GDF annotations are excluded before epoching. Downsampling to 125 Hz reduces computation and aligns with typical MI pipelines that focus on low-frequency rhythms. The 4–38 Hz band-pass filter targets mu and beta rhythms while removing slow drifts and high-frequency noise. Epoching from 3.0–7.0 seconds focuses on the main imagery period and provides a fixed-length 4 s window for all models. Per-epoch z-scoring normalizes amplitude differences across sessions and subjects without using statistics from other trials or from the test set.

E. Evaluation settings

Three evaluation regimes are defined to reflect different levels of calibration availability (Table 3). (i) Within-subject: models are trained on the subject's screening sessions (01T+02T) and tested on the subject's 03T session. This setting measures performance when the user has full calibration data available. (ii) Cross-subject LOSO: one subject is held out for testing, and models are trained on the remaining subjects' screening data (01T+02T).

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 239-262
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

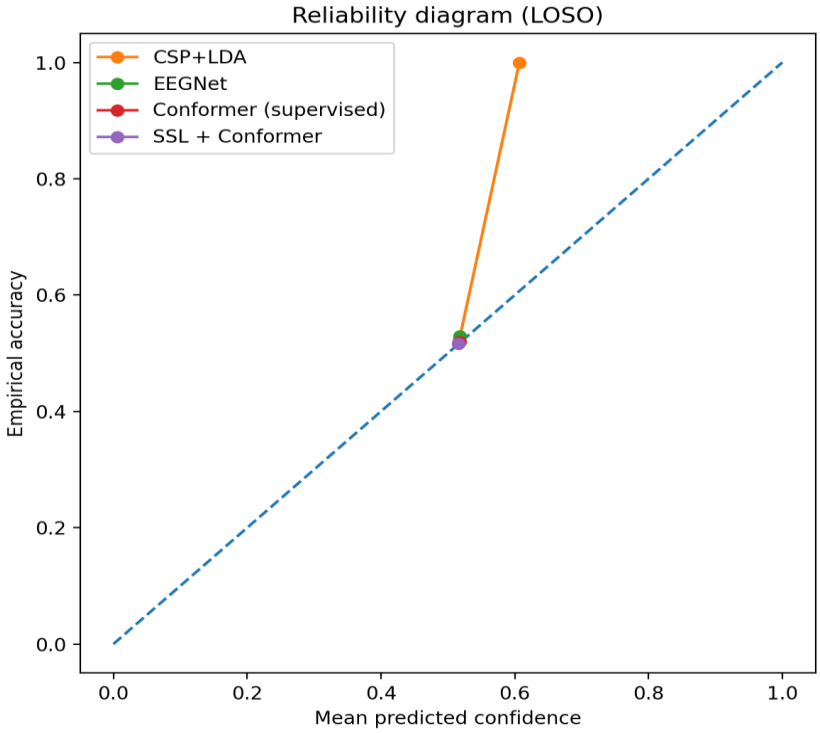

Figure 5. Reliability Diagram for Strict LOSO Evaluation (Actual Results)

Testing uses only the held-out subject's 03T session, yielding a strict subject-independent evaluation with no labeled data from the test subject. (iii) Few-shot adaptation: models are first trained on the training subjects as in LOSO, then fine-tuned using k labeled trials per class from the held-out subject's screening sessions, and tested on 03T. Few-shot adaptation evaluates the calibration-light regime where a small number of labeled trials are collected from a new user.

Neural models use a fixed stratified 80/20 validation split on the source training data. In within-subject evaluation, the split is drawn from the subject's 01T+02T trials. In LOSO evaluation, the split is drawn from the pooled source-subject data only, so the held-out subject remains completely unseen before testing. In few-shot adaptation, the source-trained checkpoint is reused and only the classification head is adapted using the sampled k trials per class from the held-out subject; session 03T is never used for model selection. All split indices are generated with a fixed random seed.

F. Baseline 1: CSP+LDA

The CSP+LDA pipeline follows standard MI practice. CSP solves a generalized eigenvalue problem to find spatial filters W that maximize the variance ratio between classes. Given class covariance matrices $\Sigma 1$ and $\Sigma 2$, CSP maximizes $(w^T\Sigma 1w)/(w^T\Sigma 2w)$ under normalization, yielding filters ordered by discriminability. We use m spatial filters from both ends of the spectrum (e.g., m=2 per class) and compute log-variance features of the projected signals. A linear discriminant

classifier is trained on these features. CSP+LDA is computationally efficient and provides an interpretable baseline (Blankertz et al., 2008; Lotte et al., 2007).

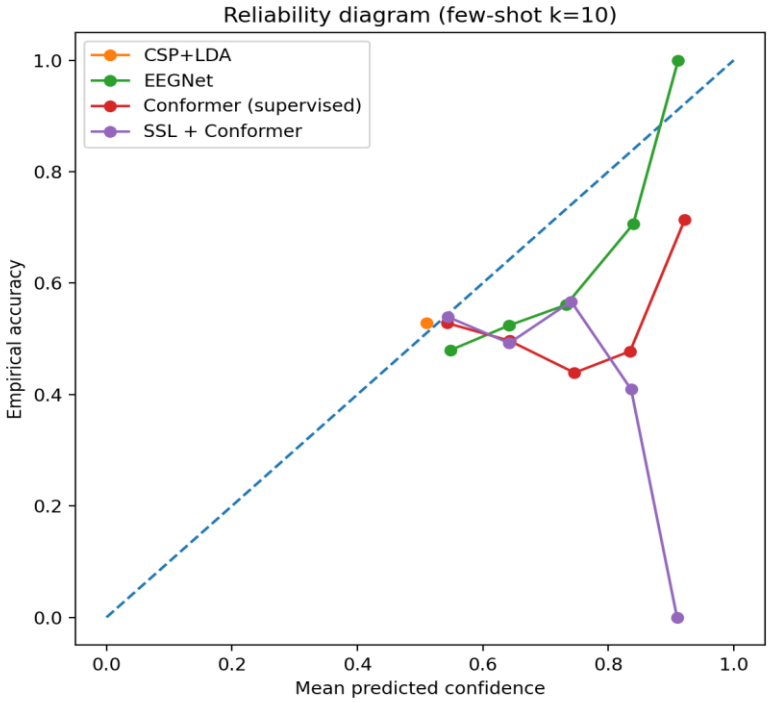

Figure 6. Reliability Diagram for Few-Shot Adaptation (k = 10) (Actual Results)

G. Baseline 2: EEGNet

EEGNet uses a temporal convolution to learn frequency-selective filters, followed by depthwise spatial convolutions to learn channel combinations, and separable convolutions to combine temporal patterns. Its compact design reduces overfitting on small datasets and is frequently used in BCI benchmarks (Lawhern et al., 2018). We specify EEGNet hyperparameters in Table 5. The learning objective is cross-entropy, and dropout is used in all convolutional blocks to reduce co-adaptation. EEGNet's efficiency also makes it suitable for latency-sensitive BCI settings, making it a strong baseline when considering real-time constraints.

H. Baseline 3: DeepConvNet and ShallowFBCSPNet

DeepConvNet is a deeper CNN with multiple convolution–pooling blocks, offering greater capacity to represent complex patterns. ShallowFBCSPNet is designed to mimic band-power pipelines by using a large temporal kernel, spatial filtering, squaring nonlinearity, log transform, and pooling. These models are widely used in EEG decoding studies and provide complementary baselines that represent high-capacity, physiologically motivated CNN designs (Schirrmeister et al., 2017). In a fair comparison, all CNNs use the same input window, the same optimization strategy family (AdamW), and comparable early stopping criteria.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 1 April 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 239-262
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

I. Baseline 4: Supervised Conformer

The supervised Conformer baseline uses the same lightweight encoder as the proposed method but is trained from scratch using labeled trials only. The encoder begins with a one-dimensional convolutional stem (stride 8, embedding dimension 32), followed by a single Conformer block with four attention heads and convolution kernel size 15. Global average pooling and a two-layer MLP generate class logits. This baseline tests whether any benefit arises from SSL pretraining rather than from the architecture alone.

J. Proposed method: SSL + Conformer

The proposed approach pretrains the Conformer encoder using masked EEG modeling. EEG epochs are divided into temporal patches, and a fixed proportion of patches is masked. The model is trained to reconstruct masked samples, using mean squared error on masked positions. After pretraining, a classification head is attached, and the entire network is fine-tuned on labeled MI trials. In LOSO evaluation, pretraining uses only training subjects to maintain a strict subject-independent setting. In few-shot adaptation, fine-tuning uses the specified k trials per class from the held-out subject.

K. Pretraining–fine-tuning interface

During pretraining, the classifier head is not used; the encoder outputs a latent sequence that is fed to a lightweight reconstruction head. During fine-tuning, the reconstruction head is removed and replaced by a classifier head that outputs class logits. The encoder parameters are initialized from the SSL checkpoint and are updated during fine-tuning. This design ensures that the downstream classifier benefits from the representation learned through reconstruction while keeping the inference-time architecture identical to the supervised Conformer baseline.

L. Optimization details

Neural models use AdamW with weight decay and early stopping. EEGNet is trained for up to 10 epochs in the within-subject setting and 6 epochs in LOSO/few-shot source training. DeepConvNet is trained for up to 8 epochs, and ShallowFBCSPNet for up to 6 epochs. The lightweight Conformer uses up to 4 pretraining epochs and 8 fine-tuning epochs in the within-subject setting, and up to 3 pretraining epochs plus 4 fine-tuning epochs in LOSO/few-shot evaluation. Fine-tuning uses cross-entropy loss, clips gradients to 1.0, and fixes all random seeds to 42.

M. Metric computation details

Accuracy is computed as the fraction of correctly classified trials. Cohen's kappa is computed as $(p_o - p_e)/(1 - p_e)$, where p_o is observed agreement and p_e is expected agreement under class marginals. Macro-F1 is computed by averaging the class-wise F1 scores, where F1 = 2·(precision·recall)/(precision+recall). ECE is computed by binning predicted confidences into equal-width bins and taking a weighted average of |acc(bin) − conf(bin)| over bins. The Brier score is the mean squared difference between predicted probability vectors and one-hot targets (Brier, 1950; Guo et al., 2017). All metrics are computed per subject and per evaluation setting to preserve interpretability of individual differences.

N. Engineering measurements

Model size is measured by serializing model parameters to disk and recording the resulting file size in megabytes. Inference latency is measured in a fixed CPU environment using repeated forward passes on single trials. To ensure comparability, all timing experiments disable gradient computation and run in evaluation mode, and warm-up passes are discarded. Latency is reported as the median and interquartile range, and the same measurement protocol is applied to every neural model to prevent biased comparisons. These measurements provide a deployment-relevant complement to decoding performance.

O. Latency measurement protocol

Inference latency is measured in single-thread CPU mode for all neural models. Each model performs 30 warm-up passes followed by 300 timed forward passes on batch size 1, and the median latency per trial together with the interquartile range is reported. This protocol isolates model computational cost and provides a deployment-relevant measure of responsiveness in an online BCI loop.

P. Statistical analysis

Comparisons are conducted across the nine subjects by treating each subject's performance as an independent sample. Paired tests compare methods within the same evaluation setting. We specify a Wilcoxon signed-rank test due to the small sample size and potential non-normality, with Holm–Bonferroni correction for multiple comparisons. We also report effect sizes (Table 8) and bootstrap confidence intervals. All statistical outputs are reported alongside per-subject scores to enable readers to inspect subject heterogeneity and to identify whether a subset of subjects drives improvements.

Q. Better-than-random criterion

Because MI datasets may have limited trial counts, raw accuracy is interpreted with caution. This study, therefore, reports per-subject accuracies together with Cohen's kappa and paired cross-

subject statistical tests, rather than relying on average accuracy alone. The better-than-random discussion of Müller-Putz et al. (2008) is used as an interpretive reference, but formal binomial chance-threshold tables are not used as the primary decision criterion in the present comparisons.

R. Ablation analysis

To isolate the contribution of SSL configuration choices, representative held-out ablations were performed on subjects B01, B04, and B08. The reported ablations vary the mask ratio (0.3/0.5/0.7) and Conformer depth (1 versus 2 blocks) while keeping the final lightweight tokenization fixed at a stride of 8. These ablations are reported as representative design diagnostics rather than as a full nine-subject benchmark.

S. Implementation and software

The pipeline is implemented in Python with reproducibility controls that fix random seeds and log all split indices. EEG preprocessing is implemented using standard digital signal processing operations (Oppenheim & Schafer, 2009). For researchers integrating with existing EEG toolchains, EEGLAB provides a common reference point for EEG analysis workflows (Delorme & Makeig, 2004). The manuscript itself is generated programmatically to ensure that figures, tables, and text remain synchronized with the method specification.

IV. RESULT

Experiments were executed on all nine subjects using the protocol defined in Section III. Trials marked with event 1023 were excluded before preprocessing, and the retained trials were filtered, downsampled, epoched, and z-scored as described above. Full within-subject benchmarking included CSP+LDA, EEGNet, DeepConvNet, ShallowFBCSPNet, supervised Conformer, and SSL+Conformer. Because the paper's main contribution concerns calibration-light transfer, the strict subject-independent and few-shot analyses focused on the four core methods: CSP+LDA, EEGNet, supervised Conformer, and SSL+Conformer.

Within-subject evaluation showed that ShallowFBCSPNet achieved the strongest full-calibration performance, with a mean accuracy of 62.23% ± 14.16%. CSP+LDA ranked second at 58.03% ± 6.70%. The proposed SSL+Conformer achieved 54.85% ± 11.15%, slightly higher than the supervised Conformer (53.56% ± 8.81%), EEGNet (52.16% ± 7.33%), and DeepConvNet (53.91% ± 11.34%), but not higher than the best shallow/classical baselines.

Under strict LOSO evaluation, average subject-independent performance was substantially lower for all methods. EEGNet achieved the highest mean accuracy at 52.92% ± 8.25%, followed by CSP+LDA at 51.76% ± 10.41%. The proposed SSL+Conformer achieved 51.56% ± 7.18%,

comparable to the supervised Conformer (51.72% ± 6.12%), but did not surpass EEGNet in the zero-calibration setting.

Few-shot adaptation showed a different pattern. At k = 1 and k = 5 trials per class, EEGNet achieved the highest mean accuracies (53.41% and 53.17%, respectively). At k = 10, however, SSL+Conformer obtained the highest mean accuracy among the core calibration-light methods at 52.84% ± 7.64%, slightly above CSP+LDA (52.40% ± 8.70%) and supervised Conformer (50.21% ± 7.99%), and clearly above EEGNet (49.12% ± 7.13%). Thus, the benefit of SSL pretraining was most visible when a modest amount of target-subject calibration data was available.

Paired Wilcoxon signed-rank tests on subject-level accuracies did not yield Holm-corrected significant differences in the within-subject or LOSO settings. In the few-shot k = 10 setting, SSL+Conformer improved over the supervised Conformer in the raw paired test ($p = 0.0195$, Cohen's $d = 0.8021$), but the difference did not remain below the 0.05 threshold after Holm correction ($p_{Holm} = 0.0586$). Against EEGNet, the raw p-value was 0.0547 (Cohen's $d = 0.5605$). Accordingly, the present results support competitive performance and a favorable trend in the k = 10 regime, but not a statistically conclusive claim of superiority.

Calibration results were mixed rather than uniformly favorable to one model. In LOSO evaluation, the supervised Conformer achieved the lowest mean ECE (0.0417), whereas SSL+Conformer achieved a slightly better mean Brier score (0.4995) than the supervised Conformer (0.4998) and CSP+LDA (0.5016). In few-shot k = 10, CSP+LDA remained the best-calibrated model by ECE (0.0667) and Brier score (0.5013), while SSL+Conformer reached ECE = 0.1313 and Brier = 0.5335.

From an engineering perspective, the proposed model remained lightweight. SSL+Conformer occupied 0.1329 MB and achieved a median single-trial CPU latency of 0.8777 ms (IQR 0.0653 ms), which was slightly faster than the supervised Conformer (0.9003 ms) and below 1 ms/trial. EEGNet remained the smallest model (0.0136 MB), whereas DeepConvNet was the largest at 1.0343 MB.

Representative ablations on held-out subjects B01, B04, and B08 showed that mask ratios 0.3 and 0.5 produced nearly identical mean accuracies (57.07% and 56.97%, respectively), whereas 0.7 was slightly worse (56.29%). Increasing Conformer depth from 1 to 2 reduced mean accuracy from 56.97% to 51.30%. Overall, the experiments show that the proposed SSL+Conformer model is a competitive calibration-light baseline, especially at k = 10, but that the strongest results on Dataset 2b remain setting-dependent and do not uniformly favor one architecture across all regimes.

V. CONCLUSION AND RECOMMENDATION

This study evaluated a calibration-light, subject-independent MI-BCI framework that combines masked-EEG modeling pretraining with a lightweight Conformer on the BCI Competition IV Dataset 2b, using only sessions 01T–03T. The experiments showed that the proposed SSL+Conformer model was competitive across settings and achieved its best relative behavior in the moderate few-shot regime. In particular, it slightly outperformed the supervised Conformer in within-subject evaluation (54.85% versus 53.56% mean accuracy) and achieved the highest mean accuracy among the core calibration-light methods at few-shot k = 10 (52.84%). The model also remained small (0.1329 MB) and fast (0.8777 ms/trial), which supports deployment feasibility.

However, the results also show that the proposed approach was not uniformly superior. ShallowFBCSPNet remained strongest in the fully calibrated within-subject setting, and EEGNet achieved the best mean accuracy under strict LOSO evaluation. In addition, no pairwise difference remained statistically significant after Holm correction across the predefined subject-level comparisons. Therefore, the present evidence supports SSL+Conformer as a competitive calibration-light benchmark rather than as a definitive replacement for classical or compact CNN baselines on Dataset 2b.

The study is further limited by the scope of Dataset 2b, which contains only three bipolar channels and two classes, and by the fact that the reported mask-ratio and depth ablations were conducted on representative held-out subjects rather than all nine subjects. Future work should validate the same protocol on additional MI datasets, improve probability calibration under few-shot adaptation, and investigate whether richer pretraining objectives or subject-conditional adaptation can convert the favorable k = 10 trend into a statistically reliable gain.

REFERENCES

Azab, A. M., Mihaylova, L., Ang, K. K., & Arvaneh, M. (2019). Weighted Transfer Learning for Improving Motor Imagery-Based Brain–Computer Interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 27*(7), 1352–1359. https://doi.org/10.1109/tnsre.2019.2923315

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Müller, K.-R. (2008). Optimizing Spatial Filters for Robust EEG Single-Trial Analysis. *IEEE Signal Processing Magazine, 25*(1), 41–56. https://doi.org/10.1109/msp.2008.4408441

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review, 78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078

Brunner, C., Leeb, R., Müller-Putz, G. R., Schlögl, A., & Pfurtscheller, G. (2008). *BCI Competition 2008 – Graz Data Set B*. Graz University of Technology. https://bbci.de/competition/iv/desc_2b.pdf

Cai, M., & Zeng, Y. (2024). MAE-EEG-Transformer: A Transformer-Based Approach Combining Masked Autoencoder and Cross-Individual Data Augmentation Pre-Training for EEG Classification. *Biomedical Signal Processing and Control, 94*, 106131. https://doi.org/10.1016/j.bspc.2024.106131

Delorme, A., & Makeig, S. (2004). EEGLAB: An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis. *Journal of Neuroscience Methods, 134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Fu, Z., Zhu, H., Zhao, Y., Huan, R., Zhang, Y., Chen, S., & Pan, Y. (2024). GMAEEG: A Self-Supervised Graph Masked Autoencoder for EEG Representation Learning. *IEEE Journal of Biomedical and Health Informatics, 28*(11), 6486–6497. https://doi.org/10.1109/jbhi.2024.3443651

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., & Wu, Y. (2020). Conformer: Convolution-Augmented Transformer for Speech Recognition. *Proceedings of Interspeech 2020*, 5036–5040. https://doi.org/10.21437/interspeech.2020-3015

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

Handoko, M., Parancika, R. B., Aris, M., & Ardi, Y. M. (2025). Determination of Employee Performance: Work Environment and Leadership Style (Case Study at PT MPIW Jakarta). *Journal of Management and Informatics (JMI)*, *4*(2), 773–790. https://doi.org/10.51903/jmi.v4i2.216

He, H., & Wu, D. (2020). Transfer Learning for Brain–Computer Interfaces: A Euclidean Space Data Alignment Approach. *IEEE Transactions on Biomedical Engineering, 67*(2), 399–410. https://doi.org/10.1109/tbme.2019.2913914

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009. https://doi.org/10.1109/cvpr52688.2022.01574

Hendry, H., & Manongga, D. (2024). Implementation of Multi-Node Sensor Data Delivery Using the Master-Slave Method in LoRa Communication. *Journal of Technology Informatics and Engineering (JTIE)*, *3*(2), 117–137. https://jtie.stekom.ac.id/index.php/jtie/article/view/279

Juan, J. V., Martínez, R., Iáñez, E., Ortiz, M., Tornero, J., & Azorín, J. M. (2024). Exploring EEG-Based Motor Imagery Decoding: A Dual Approach Using Spatial Features and Spectro-Spatial Deep Learning Model IFNet. *Frontiers in Neuroinformatics, 18*, 1345425. https://doi.org/10.3389/fninf.2024.1345425

Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6980

Ko, W., Jeon, E., Jeong, S., Phyo, J., & Suk, H.-I. (2021). A Survey on Deep Learning-Based Short/Zero-Calibration Approaches for EEG-Based Brain–Computer Interfaces. *Frontiers in Human Neuroscience, 15*, 643386. https://doi.org/10.3389/fnhum.2021.643386

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A Compact Convolutional Neural Network for EEG-Based Brain–Computer Interfaces. *Journal of Neural Engineering, 15*(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

Leeb, R., Brunner, C., Müller-Putz, G. R., Schlögl, A., & Pfurtscheller, G. (2008). BCI Competition 2008 – Graz Data Set B (BCI Competition IV Dataset 2b): Description. *Graz University of Technology*. https://bbci.de/competition/iv/desc_2b.pdf

Li, M., & Xu, D. (2024). Transfer Learning in Motor Imagery Brain Computer Interface: A Review. *Journal of Shanghai Jiaotong University (Science), 29*, 37–59. https://doi.org/10.1007/s12204-022-2488-4

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A Review of Classification Algorithms for EEG-Based Brain–Computer Interfaces. *Journal of Neural Engineering, 4*(2), 1–13. https://doi.org/10.1088/1741-2560/4/2/r01

Müller-Putz, G. R., Scherer, R., Brunner, C., Leeb, R., & Pfurtscheller, G. (2008). Better Than Random: A Closer Look on BCI Results. *International Journal of Bioelectromagnetism, 10*(1), 52–55. http://www.ijbem.org/volume10/number1/papers/paper7.pdf

Oppenheim, A. V., & Schafer, R. W. (2009). *Discrete-Time Signal Processing* (3rd ed.). Pearson. https://www.pearson.com/en-us/subject-catalog/p/discrete-time-signal-processing/P200000003144

Pfurtscheller, G., & Neuper, C. (2001). Motor Imagery and Direct Brain–Computer Communication. *Proceedings of the IEEE, 89*(7), 1123–1134. https://doi.org/10.1109/5.939829

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep Learning with Convolutional Neural Networks for EEG Decoding and Visualization. *Human Brain Mapping, 38*(11), 5391–5420. https://doi.org/10.1002/hbm.23730

She, Q., Chen, T., Fang, F., Zhang, J., Gao, Y., & Zhang, Y. (2023). Improved Domain Adaptation Network Based on Wasserstein Distance for Motor Imagery EEG Classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31*, 1137–1148. https://doi.org/10.1109/tnsre.2023.3241846

Tangermann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Miller, K. J., Müller-Putz, G., Nolte, G., Pfurtscheller, G., Preissl, H., Schalk,

G., Schlögl, A., Vidaurre, C., Waldert, S., & Blankertz, B. (2012). Review of the BCI Competition IV. *Frontiers in Neuroscience, 6*, 55. https://doi.org/10.3389/fnins.2012.00055

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 5998–6008. https://arxiv.org/abs/1706.03762

Wang, X., Liesaputra, V., Liu, Z., Wang, Y., & Huang, Z. (2024). An In-Depth Survey on Deep Learning-Based Motor Imagery Electroencephalogram (EEG) Classification. *Artificial Intelligence in Medicine, 147*, 102738. https://doi.org/10.1016/j.artmed.2023.102738

Wimpff, M., Gizzi, L., Zerfowski, J., & Yang, B. (2024). EEG Motor Imagery Decoding: A Framework for Comparative Analysis With Channel Attention Mechanisms. *Journal of Neural Engineering, 21*(3), 036020. https://doi.org/10.1088/1741-2552/ad48b9

Zainudin, A., Hadi, A. P., & Priyadi, A. (2024). Sistem Informasi Persediaan Obat Berbasis Web di Rumah Sakit Bina Kasih. *JUISI: Jurnal Ilmiah Sistem Informasi*, *3*(3), 30–34. https://doi.org/10.51903/776j7727