What Robustness Evaluation Is Needed Toward Practical Usage of EEG Decoding Models?

Jingyuan Wang
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
jw24782@essex.ac.uk

Junhua Li
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
junhua.li@essex.ac.uk

***Abstract*—Electroencephalogram (EEG)-based brain computer interfaces (BCIs) are moving toward practical deployment, robustness to channel variability becomes essential due to unstable multichannel EEG acquisition. In real-world scenarios, electrode detachment can lead to missing channels. This situation can be simulated by channel removal. In addition to missing channels, channel identity mismatch could happen, where the channel order or labels do not correctly match the default electrode layouts. This situation can be simulated by channel permutation. This study investigates the model robustness in both channel removal and channel permutation using four representative EEG decoding models. The results demonstrate that the performance decreases as the proportion of channel removal increases. Even with 50% channel removal, accuracy remains above 0.29, showing that the model can still maintain reasonable performance under substantial channel removal. In contrast, channel permutation results in a sharp performance degradation to near-chance levels, with accuracy falling to approximately 0.262-0.293 across all models. Although both cases impair performance, the models are considerably more sensitive to channel permutation. These findings suggest that practical deployment of EEG-based BCI should assess robustness to both missing channels and channel-order variability, alongside conventional classification performance metrics.**

***Index Terms*—brain computer interfaces (BCIs), robustness evaluation, EEG, deep learning, channel permutation.**

I. Introduction

Deep learning models are increasingly used in EEG-based BCIs to decode neural activity in tasks such as motor imagery classification [1]–[5]. As decoding performance approaches practical usability, robustness has emerged as a primary bottleneck for real-world EEG-based BCI deployment. In real-world settings, sensor availability, electrode quality, and recording conditions often vary across subjects and experimental sessions [6], [7]. These conditions make it essential to systematically evaluate model robustness to channel-related variability. One of the evaluating robustness is channel removal, where a subset of channels is removed and the resulting performance degradation is reported. This practice is often used to examine whether a model depends heavily on specific channels and whether it can tolerate channel missing due to noise or hardware failure. However, channel removal evaluates robustness only under a fixed channel layouts, where the correspondence between input channels and physiological electrode layouts is assumed to remain unchanged. In practical EEG-based BCI systems, EEG deployment may also face channel identity mismatch. In this case, all EEG signals may be available, but the channel order or channel labels may not match the default electrode layouts. This may occur due to device-dependent electrode layouts, changed channel ordering during preprocessing, or missing or incorrectly recorded electrodes information. In such cases, signal availability is preserved while channel identity is not. Therefore, both channel removal and channel permutation are necessary for evaluating EEG model robustness in realistic EEG-based BCI deployment.

Deep learning models for EEG decoding span a wide range of modeling paradigms, including spatial filtering inspired networks, convolutional neural networks, and more recent attention-based models [1], [8], [9]. These models differ substantially in how they aggregate information across channels and time, yet they are often evaluated using similar robustness assessment protocols. At the same time, recent studies have highlighted the importance of cross-subject and cross-session evaluation for assessing generalization in EEG decoding [10]–[13]. Compared with within-subject settings, these protocols reduce subject-specific overfitting and better reflect practical EEG-based BCI deployment, where frequent retraining is often infeasible. However, practical EEG-based BCI models are typically trained and applied on heterogeneous data spanning multiple users and recording sessions, rather than being evaluated on a specific subject or session. So multi-subject multi-session evaluation protocol is adopted in this study to better reflect practical EEG-based BCI deployment.

To address channel variability relevant to the deployment in EEG-based BCIs, robustness must be evaluated under channel removal and channel permutation that reflect realistic acquisition conditions. Channel removal simulates electrode detachment or signal loss, whereas channel permutation simulates channel identity mismatch. Two strategies in channel permutation are examined, including global permutation and per-trial permutation. Model performance under these perturbations is evaluated in multi-subject multi-session settings to analyze the relative impact of channel missing and channel identity mismatch. The objective is to establish a robustness evaluation scheme that aligns with practical EEG-based BCI deployment requirements. Specifically, the main contributions are summarized as follows:

TABLE I
EEG DECODING MODELS EVALUATED IN THIS STUDY.

Model	Model Paradigm
ShallowFBCSPNet	CSP-Inspired spatial filtering
Deep4Net	Deep convolutional neural network
EEGNet	Compact CNN with depthwise convolutionals
EEGConformer	Hybrid CNN-Transformer with self-attention

- Evaluation of model robustness under channel removal and channel permutation to reflect realistic channel variability in practical EEG-based BCI systems.
- Implementation of two permutation strategies (i.e., global and per-trial) to simulate channel order inconsistencies.
- Comparative analysis of four representative EEG decoding models, including spatial-filtering-based, convolutional-based, and attention-based models, in a multi-subject multi-session setting.

II. METHOD

A. Dataset and Evaluation Setting

We conduct experiments on a publicly available motor imagery EEG dataset BNCI2014_001 from the MOABB benchmark framework [14]. This dataset involved four-class motor imagery tasks (left hand, right hand, both feet, and tongue). Nine subjects participated, with each subject completing two sessions on different days. Each session comprised six runs, separated by brief breaks. Within each run, 48 trials were performed (12 trials per class), yielding a total of 288 trials per session. EEG data were acquired using 22 active Ag/AgCl electrodes placed according to the international 10–20 system, at a sampling rate of 250 Hz.

To approximate practical EEG-based BCI deployment, a multi-subject multi-session evaluation protocol is adopted. Data from all subjects and recording sessions are combined and randomly partitioned into training and testing sets, without sharing trials across splits. This setting exposes models to inter-subject and inter-session variability during both training and evaluation, rather than restricting learning to subject-specific or session-specific distributions. Compared with within-subject or within-session evaluation, this protocol reduces overfitting to individual recording characteristics and better reflects realistic EEG-based BCI scenarios. Moreover, variability across subjects and sessions introduces differences in channel characteristics and spatial configurations, making this setting suitable for analyzing dependence on channel identity and structural organization.

B. Models

We evaluate representative EEG decoding models covering different modeling paradigms. These models are widely used in the literature and serve as standard baselines rather than proposed methods. A comparison of the four EEG decoding models is summarized in Table I.

All models are used with their default architectures as provided by the Braindecode library. No model-specific architectural modifications or hyperparameter tuning are performed to ensure fair comparison.

C. Robustness Probing Protocol

1) Channel Removal: Channel Removal is a commonly used robustness evaluation method. Formally, channel removal can be expressed as $\tilde{\mathbf{X}} = \mathbf{M} \odot \mathbf{X}$, where $\mathbf{M} \in \{0,1\}^{C\times 1}$ is a binary channel mask. During testing, a fixed proportion p of channels is removed, and the resulting performance degradation is calculated. This procedure simulates channel missing caused by sensor failure or severe noise. In this work, channel removal serves as a baseline robustness test. We evaluate three multiple removal ratios p 10%, 30%, 50% and report average performance over repeated random channel selections.

2) Channel Permutation: Channel permutation does not only aim to simulate a specific real-world failure mode, but also serves as a stress test that isolates dependence on channel identity while preserving all signal values. To evaluate robustness to channel identity, we introduce channel permutation. Channel permutation is defined as $\tilde{\mathbf{X}} = \mathbf{P}\mathbf{X}$, where $\mathbf{P} \in \{0,1\}^{C\times C}$ is a permutation matrix. This operation preserves all signal values while altering channel identity. Unlike channel removal, permutation alters the mapping between signals and channel indices while preserving all signal values. We consider two permutation settings:

- **Global permutation**, where the same channel permutation is applied to all trials.
- **Per-trial permutation**, where each trial is permuted independently.

These perturbations preserve signal data while permuting the order of channel. Comparing model behavior under channel removal and channel permutation allows us to disentangle robustness to missing channels from dependence on fixed channel layouts. If a model is invariant to channel identity, its performance should remain stable under channel permutation.

3) Experimental Protocol: All experiments follow a unified and reproducible protocol. We employ 5-fold stratified cross-validation and repeat experiments using three different random seeds 42, 123 and 256. Models are trained for a maximum of 100 epochs with early stopping based on validation performance. EEG signals are band-pass filtered between 0.5–45 Hz and notch-filtered at 50 Hz. Z-score normalization is applied using statistics computed from the training data only and then applied to test data to avoid information leakage. Model performance is evaluated using classification accuracy, Cohen's kappa, and macro-averaged F1 score. Accuracy reflects overall classification performance, while Cohen's kappa accounts for chance-level agreement in multi-class settings. Macro-averaged F1 score is used to mitigate class imbalance effects. No model-specific hyperparameter tuning is performed.

TABLE II

NO PERMUTATION PERFORMANCE AND PERMUTATION PROBING ON BNCI2014_001 (4-CLASS). MEAN±STD OVER 15 RUNS PER MODEL (3 RANDOM SEEDS × 5 FOLDS). CHANNEL IDENTITY SENSITIVITY (CIS) IS THE ACCURACY DIFFERENCE FROM NO PERMUTATION TO PERMUTED.

Model	**No Permutation**			**Global Permutation**		**Per-Trial Permutation**		**CIS (Accuracy Difference)**	
	Accuracy	Kappa	$F1_{macro}$	Accuracy	Kappa	Accuracy	Kappa	Global	Per-Trial
Deep4Net	0.614±0.019	0.485±0.025	0.611±0.018	0.287±0.025	0.049±0.033	0.293±0.011	0.058±0.015	0.327±0.036	0.320±0.026
EEGConformer	0.605±0.027	0.474±0.036	0.604±0.026	0.262±0.017	0.015±0.023	0.277±0.008	0.037±0.011	0.344±0.034	0.328±0.027
EEGNet	0.591±0.022	0.455±0.029	0.590±0.022	0.277±0.032	0.036±0.042	0.277±0.013	0.036±0.018	0.314±0.035	0.314±0.026
ShallowFBCSPNet	0.600±0.028	0.466±0.037	0.599±0.028	0.271±0.016	0.027±0.022	0.277±0.013	0.036±0.017	0.329±0.037	0.323±0.032

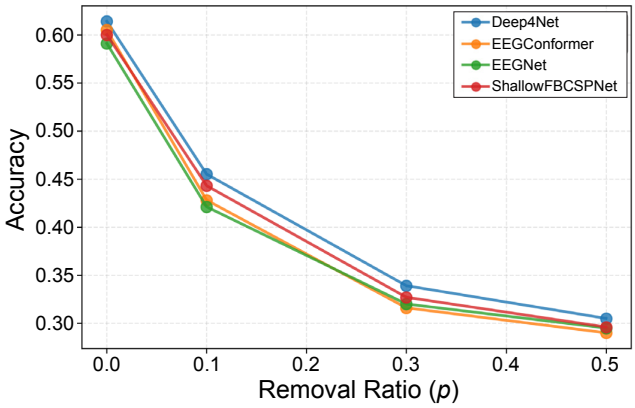

Fig. 1. Classification accuracy under increasing channel removal ratios. Random channel removal lead to gradual performance degradation across all evaluated models, suggesting apparent robustness to channel missing.

III. RESULTS AND DISCUSSION

All models achieved better performance than chance-level on BNCI2014-001 under the multi-subject multi-session protocol, indicating that the task remains learnable in a deployment-oriented setting. Table II reports performance under the no-removal and no-permutation condition. To reduce the influence of random initialization, each model was trained using three different random seeds within a five-fold cross-validation scheme, resulting in 15 runs per model (3 seeds × 5 folds). Reported results correspond to the mean and standard deviation across these runs. Deep4Net obtained the highest average accuracy (0.614±0.019) and kappa (0.485±0.025). EEGConformer and ShallowFBCSPNet achieved comparable accuracy (0.605±0.027 and 0.600±0.028), while EEGNet showed slightly lower performance (0.591±0.022).

We first analyze robustness using channel removal protocol. Figure 1 illustrates accuracy as a function of removal probability for random channel removal. Across all models, accuracy decreased smoothly as the removal ratio increased, and performance remained clearly above chance even when half of the channels were removed. The overall shapes of the degradation curves were similar across models, suggesting that none of the models relied exclusively on a small subset of electrodes. From this perspective alone, all models would be considered reasonably robust to channel missing. However, this conclusion critically depends on evaluating robustness only through channel removal.

Furthermore, as shown in Table III, random channel removal leads to a consistent decrease in accuracy across four models. When $p = 0.1$, the accuracy of all models decreases by approximately 0.157–0.177 compared with the no channel removal setting, showing that even the loss of a small number of channels can substantially affect model performance. When the channel removal ratio increases to 0.3, the accuracy of all models drops to approximately 0.316–0.339. At a channel removal ratio of 0.5, the accuracy of the all models further decreases to a lower range of 0.290–0.305. The relatively small standard deviations in accuracy indicate that this trend is stable across different random seeds. These results suggest that the evaluated EEG models do not rely solely on a few specific channels. Instead, they depend on the full set of channels. Therefore, randomly removing channels weakens the spatial information available to the models, leading to a substantial performance degradation. This finding further supports the need to develop EEG decoding methods that can better handle incomplete or inconsistent channel configurations.

A fundamentally different picture emerges when channel identity is perturbed without removing any channels. Figure 2 compares no permutation performance with global and per-trial channel permutation. In both cases, all models suffered a dramatic and consistent collapse. Accuracy dropped from approximately 0.591–0.614 under no permutation conditions to near chance level (0.262–0.293), with kappa approaching zero. This occurred despite the fact that all EEG signals were preserved and only their channel ordering was altered. The resulting channel identity sensitivity (CIS) was large and stable across models, ranging from 0.314 to 0.344 for global permutation and from 0.314 to 0.328 for per-trial permutation (Table II).

Notably, global and per-trial permutation produced very similar degradation. This suggests that model failure is dominated by breaking the learned association between channel identity and fixed electrode layouts, rather than by random permutations across trials. In other words, once channel identity is disrupted, performance collapses regardless of whether the permutation is consistent or varies across samples.

Taken together, Figures 1 and 2 show that robustness to channel removal does not necessarily imply robustness to channel perturbation. The models retain a degree of robustness to channel removal, even under substantial channel missing, whereas permutation probing reveals high sensitivity to channel identity. These results demonstrate that robustness to missing channels does not imply robustness to channel reordering or remapping. Consequently, channel removal robustness

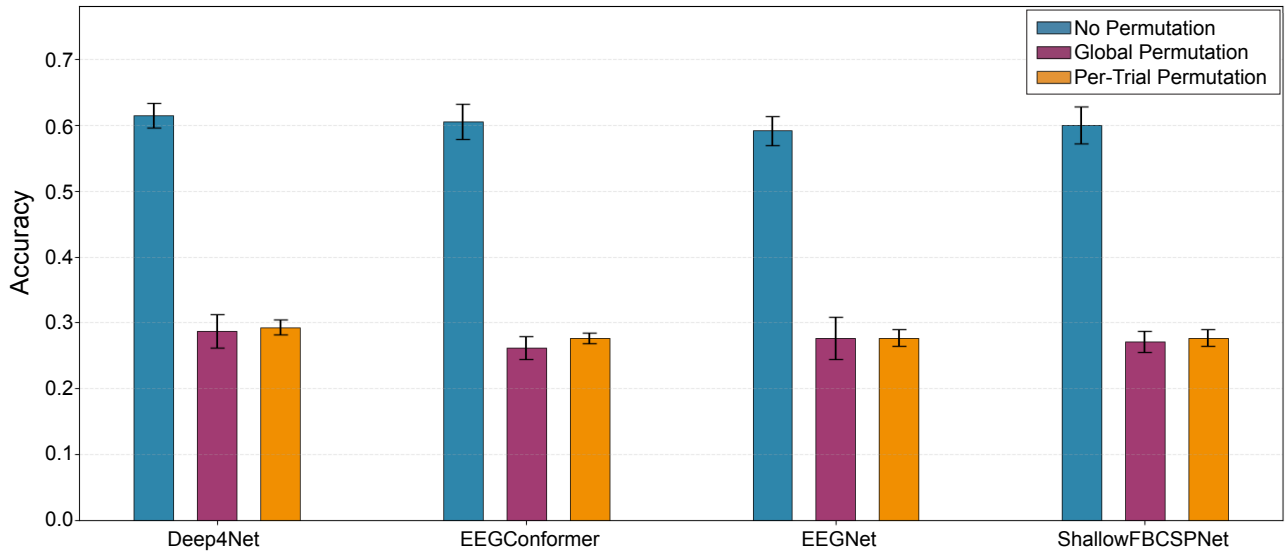

Fig. 2. Model performance under channel permutation. While all signals are preserved, both global and per-trial channel permutation cause accuracy and kappa to collapse toward chance level, indicating high dependence on fixed channel identity.

evaluations alone are insufficient to characterize how EEG decoding models behave under realistic deployment variability.

Finally, attention-based EEGConformer did not show improved robustness to channel permutation compared with convolutional or spatial-filtering-based models. Its CIS values were among the highest across models. This underscores that channel identity consistency remains a critical constraint for real-world EEG-based BCI systems, even for recently models. Overall, robustness evaluation should therefore assess both channel removal and channel permutation, since they reflect distinct deployment-relevant channel variations. Testing under both conditions offers a clearer indication of whether a model is suitable for practical EEG-based BCI deployment.

TABLE III
CHANNEL REMOVAL PROBING (ACCURACY). MEAN±STD OVER 3 RANDOM SEEDS (EACH SEED AGGREGATED OVER 5 FOLDS).

Model	p	Accuracy
Deep4Net	0.0	0.614±0.007
	0.1	0.455±0.003
	0.3	0.339±0.001
	0.5	0.305±0.003
EEGConformer	0.0	0.605±0.006
	0.1	0.428±0.003
	0.3	0.316±0.002
	0.5	0.290±0.002
EEGNet	0.0	0.591±0.010
	0.1	0.421±0.003
	0.3	0.320±0.006
	0.5	0.295±0.006
ShallowFBCSPNet	0.0	0.600±0.006
	0.1	0.443±0.004
	0.3	0.327±0.001
	0.5	0.296±0.003

IV. CONCLUSION

This study evaluates robustness of EEG decoding models under two deployment-relevant channel variations which are channel removal and channel permutation. Both cases degrade model performance, confirming that channel variability has a crucial impact on decoding performance. Performance decreases progressively as more channels are removed, indicating limited but observable tolerance to signal absence. In contrast, channel permutation consistently produces a much larger performance decline, often approaching chance level despite preserving all signal information. These results demonstrate that EEG models are substantially more sensitive to channel-order variation than to channel missing. Importantly, robustness observed under channel removal does not guarantee stability under realistic channel-order inconsistencies. A model that performs well in offline evaluation may still suffer substantial performance degradation under complex real-world deployment conditions. Therefore, practical EEG-based BCI deployment requires explicit assessment of both signal-absence robustness and channel-order robustness. Incorporating both evaluation conditions can provide a more reliable estimate of real-world model stability and help prevent unexpected performance collapse after deployment.

REFERENCES

[1] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG Conformer: Convolutional Transformer for EEG Decoding and Visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9991178

[2] G. Liu, R. Zhang, L. Tian, and W. Zhou, "Fine-Grained Spatial-Frequency-Time Framework for Motor Imagery Brain-Computer Interface," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–13, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10857659

[3] P. Autthasan, R. Chaisaen, T. Sudhawiyangkul, P. Rangpong, S. Kiatthaveephong, N. Dilokthanakul, G. Bhakdisongkhram, H. Phan, C. Guan, and T. Wilaiprasitporn, "MIN2Net: End-to-End Multi-Task Learning for Subject-Independent Motor Imagery EEG Classification," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 6, pp. 2105–2118, Jun. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9658165

[4] W.-Y. Hsu and Y.-W. Cheng, "EEG-Channel-Temporal-Spectral-Attention Correlation for Motor Imagery EEG Classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1659–1669, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10065454

[5] J. LI, J. LIANG, Q. ZHAO, J. LI, K. HONG, and L. ZHANG, "DESIGN OF ASSISTIVE WHEELCHAIR SYSTEM DIRECTLY STEERED BY HUMAN THOUGHTS," *International Journal of Neural Systems*, vol. 23, no. 03, p. 1350013, 2013, _eprint: https://doi.org/10.1142/S0129065713500135. [Online]. Available: https://doi.org/10.1142/S0129065713500135

[6] R. A. Miranda, W. D. Casebeer, A. M. Hein, J. W. Judy, E. P. Krotkov, T. L. Laabs, J. E. Manzo, K. G. Pankratz, G. A. Pratt, J. C. Sanchez, D. J. Weber, T. L. Wheeler, and G. S. Ling, "DARPA-funded efforts in the development of novel brain–computer interface technologies," *Journal of Neuroscience Methods*, vol. 244, pp. 52–67, Apr. 2015. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165027014002702

[7] H. Zhang, T. Zuo, Z. Chen, X. Wang, and P. Z. Sun, "Evolutionary Ensemble Learning for EEG-Based Cross-Subject Emotion Recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 7, pp. 3872–3881, Jul. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10490105

[8] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, Oct. 2018. [Online]. Available: https://iopscience.iop.org/article/10.1088/1741-2552/aace8c

[9] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, Nov. 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/hbm.23730

[10] W. Sun and J. Li, "AdaptEEG: A Deep Subdomain Adaptation Network With Class Confusion Loss for Cross-Subject Mental Workload Classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 3, pp. 1940–1949, Mar. 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10787106

[11] X. Li, D. Song, P. Zhang, Y. Zhang, Y. Hou, and B. Hu, "Exploring EEG Features in Cross-Subject Emotion Recognition," *Frontiers in Neuroscience*, vol. 12, p. 162, Mar. 2018. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2018.00162

[12] Y. Peng, H. Liu, J. Li, J. Huang, B.-L. Lu, and W. Kong, "Cross-Session Emotion Recognition by Joint Label-Common and Label-Specific EEG Features Exploration," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 759–768, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10003248

[13] P. Yu, X. He, H. Li, H. Dou, Y. Tan, H. Wu, and B. Chen, "FMLAN: A novel framework for cross-subject and cross-session EEG emotion recognition," *Biomedical Signal Processing and Control*, vol. 100, p. 106912, Feb. 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1746809424009704

[14] M. Tangermann, K.-R. Müller, A. Aertsen, N. Birbaumer, C. Braun, C. Brunner, R. Leeb, C. Mehring, K. J. Miller, G. R. Müller-Putz, G. Nolte, G. Pfurtscheller, H. Preissl, G. Schalk, A. Schlögl, C. Vidaurre, S. Waldert, and B. Blankertz, "Review of the BCI Competition IV," *Frontiers in Neuroscience*, vol. 6, 2012. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2012.00055