

Evaluation Leakage Creates 41 Points of False Accuracy in Brain-Computer Interfaces for Healthy and Post-Stroke Users

Aria Chadha



Code & data
on GitHub

INTRODUCTION

The Scale of the Problem

12.2M

new strokes
worldwide each year

133

participants (87
healthy, 46 stroke)

4,834

motor-imagery trials

41.3 pp

accuracy created by
leakage

Stroke is one of the leading causes of long-term disability worldwide. Many survivors lose reliable control of an arm or hand, making tasks like steering a powered wheelchair difficult. A **brain-computer interface (BCI)** offers another route, reading brain activity directly and turning it into commands. BCIs that work well for healthy volunteers often perform worse for patients, whose brain signals have been changed by the injury.

EEG and Brain-Computer Interface

- A BCI translates brain signals into **commands for external devices**.
- In assistive robotics, EEG-based BCIs let users **control powered wheelchairs**, giving mobility and independence to people with severe motor impairment.
- EEG is **safe, portable and inexpensive**, but its signals are weak and noisy, so how a decoder is built and tested matters a great deal.

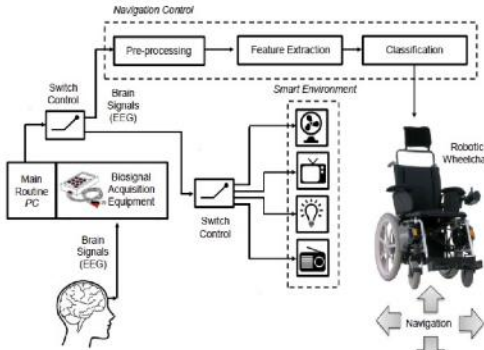


Figure 1. An example EEG-controlled wheelchair. Brain signals are pre-processed, converted to features and classified into movement commands (image source to be added).

When a person imagines moving one hand, the 8–30 Hz rhythms over the **motor cortex on the opposite side** of the brain weaken. Electrodes C3 and C4 sit directly over these areas.

The standard way to extract this pattern is **Common Spatial Patterns (CSP)**, which learns the electrode combinations that best separate left from right, using the trial labels.

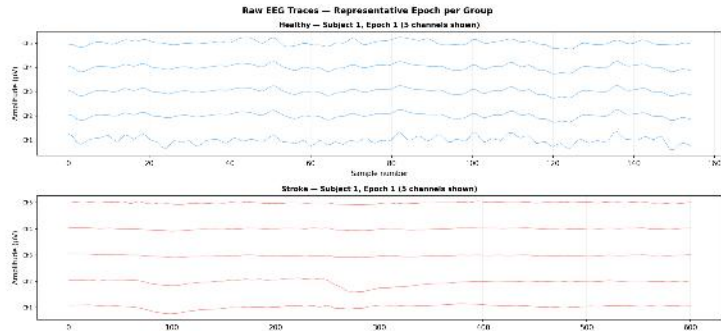


Figure 2. Representative epochs from one healthy and one stroke participant. Stroke EEG is flatter, leaving less signal to decode.

RELATED WORKS

Prior Work

Motor-imagery BCIs detect the drop in 8–30 Hz power over the sensorimotor cortex during imagined movement (Pfurtscheller & Lopes da Silva, 1999). **CSP** is the most widely used way to extract it (Ramoser et al., 2000; Blankertz et al., 2008), and Riemannian classifiers are a strong alternative (Lotte et al., 2018). EEG wheelchairs have mostly been tested with healthy users (Al-qaysi et al., 2018).

Gap in Current Research

- If CSP is fitted before the train/test split, the test trials shape the model, a form of **data leakage**. It affects hundreds of machine-learning papers across 17 fields (Kapoor & Narayanan, 2023), yet its cost in BCI research is **rarely measured**.
- Stroke patients, who would benefit most, are **underrepresented**, and fixes for their weaker signals are seldom tested with leak-free evaluation.

What My Research Adds

- A direct measurement of CSP leakage on data with no real signal and on real EEG
- A healthy vs. stroke comparison under correct evaluation
- A rigorous test of adaptive amplification and why it fails

RESEARCH OBJECTIVE

This project measures how much accuracy CSP leakage creates, and then uses leak-free evaluation to test whether adaptive signal amplification can help stroke patients. The aim is not a new decoder but an exact measurement of a common mistake, followed by a fair test of one proposed fix.

Hypothesis 1. Fitting CSP before the train/test split will inflate accuracy, most strongly when training and testing on the same person.

Hypothesis 2. Adaptive amplification of weak EEG will improve decoding accuracy for stroke patients.

METHODOLOGY

Data and Preprocessing

Group	Source	Rate	Subjects	Trials
Healthy	PhysioNet EEGMMIDB	160 Hz	87	3,076
Stroke	figshare 21679035	500 Hz	46	1,758

Public, de-identified data; each trial is imagined left- or right-hand movement (chance = 50%). Both sets were limited to their 29 shared electrodes and re-referenced to the common average, which was essential because the stroke data's original reference made signals about 7.6× smaller at Cz. Trials above 400 µV peak-to-peak were rejected.

Decoding Pipeline



CSP filters make left and right trials as different as possible; their log-variance feeds a Random Forest. A Riemannian classifier (pyRiemann) served as an independent check.

Evaluation Design

Within-subject	Cross-subject
Train and test on the same person, as when a user calibrates their own BCI.	Train on other people and test on someone new; folds grouped by participant.



Null Experiment

Both protocols were run on synthetic EEG with **no class information**, where a fair test must score about 50%, so anything higher is created by the evaluation. Each of the four conditions was repeated 20 times with a fixed seed, so results reproduce exactly from the public code.

RESULTS

Testing Hypothesis 1 on Data with No Real Signal

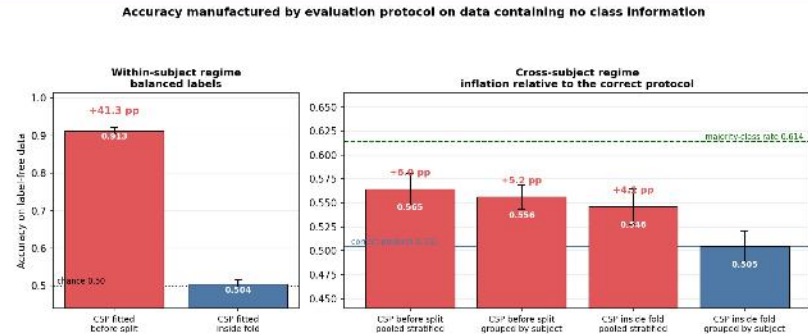


Figure 3. Accuracy on label-free data under each protocol. Created by Aria Chadha using Python Library Matplotlib.

With CSP fitted inside each training fold, accuracy was **50.4%**, as expected. Fitting it before the split gave **91.3%**, **41.3 percentage points** of fake accuracy. Across subjects it was smaller (+4.1 to +6.0 pp). In other words, a model that learned nothing would have been reported as over 90% accurate. **Hypothesis 1 was supported.**

Hypothesis 1 on Real EEG

Within-subject	Correct	Leaky	Inflation
Healthy	62.3%	76.6%	+14.3 pp
Stroke	54.1%	64.7%	+10.6 pp

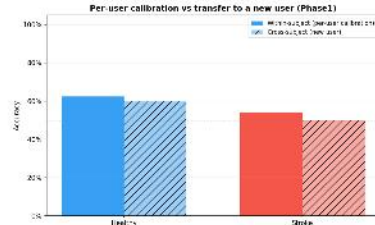


Figure 4. Leak-free accuracy for a calibrated user vs. a new user. Created by Aria Chadha using Python Library Matplotlib.

On real recordings, leakage added 10–14 points, more than the honest gap between the two groups. The leaky healthy score (76.6%) would look competitive with published work; the honest one is 62.3%. With correct evaluation, stroke accuracy was above a 200-permutation null ($p < 0.005$), but it fell to chance for a new patient (50.1% cross-subject).

Why Stroke EEG Is Harder to Decode

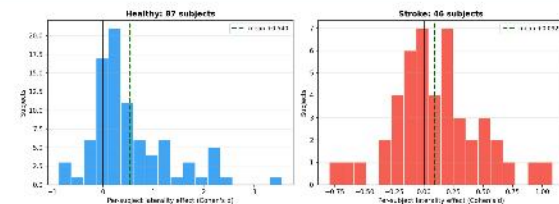


Figure 5. Per-subject left/right difference at C3 vs. C4. Created by Aria Chadha using Python Library Matplotlib.

The usual left/right difference was much weaker after stroke (mean $d = +0.54$ healthy vs. $+0.09$ stroke). Most healthy participants showed a clear shift between hemispheres; many stroke patients showed almost none.

Wheelchair Simulation

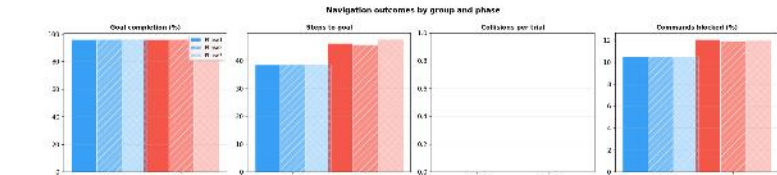


Figure 6. Simulated navigation outcomes for each group and phase. Created by Aria Chadha using Python Library Matplotlib.

Decoded commands drove a simulated wheelchair with a safety override. Both groups reached the goal in 96% of runs with zero collisions, but stroke users needed more steps (46.2 vs. 38.6) and had more commands blocked (about 12% vs. 10.5%).

TESTING HYPOTHESIS 2

Weak stroke epochs were boosted 1.1–1.5× before classification, first against a population reference, then a per-subject reference. The population reference boosted 97.9% of trials, acting as a near-constant gain, which is why the per-subject version was added. Healthy data was never amplified, so healthy accuracy is unchanged by design. Effects were tested with paired comparisons across participants, with 95% confidence intervals and Holm–Bonferroni correction.

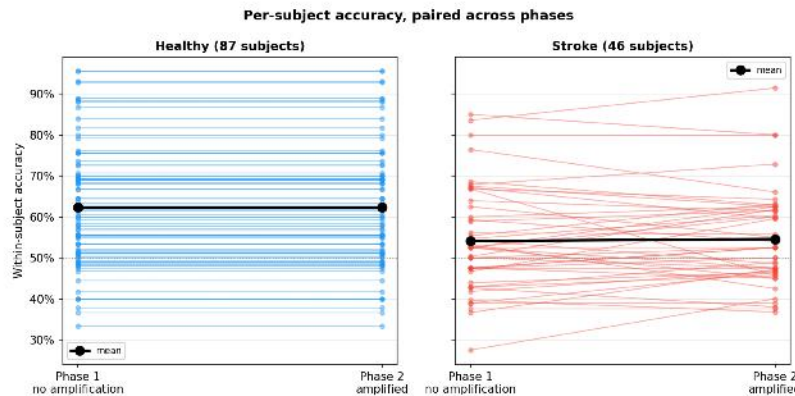


Figure 7. Each line is one participant, before and after amplification. Created by Aria Chadha using Python Library Matplotlib.

Amplification factors, subject reference

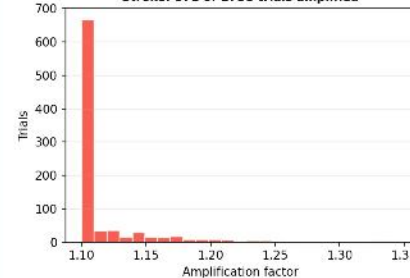


Figure 8. Amplification factors applied.

Amplification was applied to about half of all stroke trials, yet accuracy did not change. The change was **+0.72 pp** with the population reference (95% CI -1.36 to $+2.81$, $p = 0.49$) and **−0.74 pp** with the subject reference (95% CI -2.99 to $+1.52$, $p = 0.51$). A Riemannian classifier gave the same result. **Hypothesis 2 was not supported.**

DISCUSSION

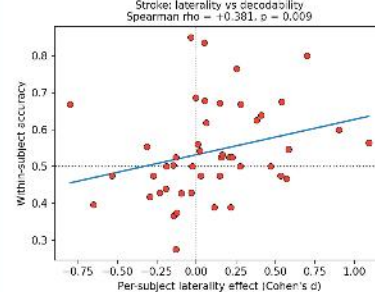


Figure 9. Stronger left/right difference, better decoding. Created by Aria Chadha using Python Library Matplotlib.

Leakage was largest within-subject, which is how real users' BCIs are calibrated. Amplification scales the whole signal, so it cannot restore the left/right difference that stroke weakens. For stroke users, a decoder may need to look beyond the usual C3/C4 pattern, for example toward the undamaged hemisphere. My earlier version of this project reported a large amplification effect, and correcting the evaluation removed it, which led to this study.

- This study used **offline recordings**, one stroke dataset and one main decoding pipeline, so results should be confirmed with live users.
- Next steps are to **audit published BCI studies** for leakage, test with live EEG hardware, and build decoders tailored to each patient.

CONCLUSION

- Fitting CSP before the split can turn random data into a 91%-accurate “BCI.”
- Leakage on real EEG (10–14 pp) is larger than the difference between healthy and stroke users.
- Signal amplification does not improve stroke decoding; stroke-specific decoders are needed.
- Every supervised step should be fitted inside the training fold, and cross-subject results should be reported alongside within-subject ones.

REFERENCES

- Al-qaysi, Z. T., Zaidan, B. B., Zaidan, A. A., & Suzani, M. S. (2018). A review of disability EEG based wheelchair control system: Coherent taxonomy, open challenges and recommendations. *Computer Methods and Programs in Biomedicine*, 164, 221–237.
- Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2012). Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4), 920–928.
- 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.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Chadha, A. (2026). Quantifying evaluation leakage in motor imagery decoding: 41 percentage points of manufactured within-subject accuracy, with a post-stroke application. *bioRxiv* 747428.
- Feigin, V. L., et al. (2022). World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. *International Journal of Stroke*, 17(1), 18–29.
- Goldberger, A. L., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23), e215–e220.
- Gramfort, A., et al. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 7, 267.
- Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804.
- Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 15.
- Lotte, F., et al. (2018). A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15(3), 031005.
- Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pfurtscheller, G., & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology*, 110(11), 1842–1857.
- Post-stroke motor imagery EEG dataset (2023). *figshare*, dataset 21679035.
- Ramoser, H., Müller-Gerking, J., & Pfurtscheller, G. (2000). Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4), 441–446.
- Schal, purp 1034.