What is the best model for decoding neurophysiological signals? Depends on how you evaluate

Bruno Aristimunha, Thomas Moreau, Sylvain Chevallier, Raphael Y de Camargo, Marie-Constance Corsi

What is the best model for decoding neurophysiological signals? Depends on how you evaluate

Bruno Aristimunha[1,2], Thomas Moreau[3], Sylvain Chevallier[1], Raphael Y. de Camargo[2], Marie-Constance Corsi[4]

1. Inria TAU, LISN-CNRS, Université Paris-Saclay, 91405, Orsay, France
2. Center for Mathematics, Computing and Cognition, Universidade Federal do ABC, Santo André, Brazil
3. Inria Mind team, Université Paris-Saclay, CEA, Palaiseau, 91120, France
4. Sorbonne Université, Institut du Cerveau – Paris Brain Institute -ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France
5. Inria NERV team, Paris, France

Non-invasive brain-computer interface (BCI) is a framework that establishes direct communication between a computational external device and the brain activity, mostly via electroencephalography (EEG) signals. Despite its clinical applications, EEG-based BCI presents several challenges, such as performance variability across subjects and a low amount of data. To tackle these issues, many approaches have been proposed to better highlight and understand the neural dynamics reflected in the signals, including signal processing tools (e.g. band-pass filters, alignments), alternative features such as functional connectivity (Corsi M.-C. *et al.* 2022), or more sophisticated classification models (based on manifolds or deep learning). Despite all these efforts, many results fail to provide a consistent answer to which type of model is the best to understand brain dynamics; even when they use the same data, they use different evaluation schemes. In this study, we are interested in the following questions: (i) Does the way the model is evaluated impact the ranking of the best model? (ii) Does the amount of data impact the decoding of the brain signals, and is this reflected in the ranking? (iii) Do the best models also deliver better interpretability? Here, we systematically evaluated different methods, 7 using deep learning, 10 using Riemannian Manifold, and Common Spatial Patterns across six EEG-based BCI datasets during the sensory-motor rhythms tasks. All these methods were benchmarked using the same data split, with the classification task determining which motor imagery task occurred during a trial. The results were consistent with prior reports (Chevallier, S. *et al.* 2024). For instance, the best deep learning model, Attention Net (Wimpff, M., Gizzi, *et al.* 2024), outperformed the best Riemannian model, Fucone (Corsi M.-C. *et al.* 2022), by 16% in five-fold cross-validation at the subject level on the BCI 2014 competition dataset. However, the ranking remained inconsistent when we changed the evaluation method. The amount of data

used as input in the model was decisive for deep learning models, while the manifold models proved to be more invariant to this factor when trained with one model for a subject. Finally, the inherent interpretability of the functional connectivity models was not effective for scenarios with many subjects. These results emphasize the necessity of a systematic comparison of brain decoding models, drawing a parallel with the benchmark approaches that built the foundation in deep learning fields, which now could be adopted in neuroscience.

References

Corsi, M. C., Chevallier, S., De Vico Fallani, F., & Yger, F. (2022). Functional connectivity ensemble method to enhance BCI performance (FUCONE). *IEEE Transactions on Biomedical Engineering*, *69*(9), 2826-2838.

Wimpff, M., Gizzi, L., Zerfowski, J. and Yang, B., 2024. EEG motor imagery decoding: A framework for comparative analysis with channel attention mechanisms. *Journal of Neural Engineering*, *21*(3), p.036020.

Chevallier, S., Carrara, I., Aristimunha, B., Guetschel, P., Sedlar, S., Lopes, B., ... & Moreau, T. (2024). The largest EEG-based BCI reproducibility study for open science: the MOABB benchmark. *arXiv preprint arXiv:2404.15319*.

Acknowledgments

The work of BA was supported by DATAIA Convergence Institute as part of the ``Programme d'Investissement d'Avenir'', (ANR-17-CONV-0003) operated by LISN.

E-mails
b.aristimunha@gmail.com
thomas.moreau@inria.fr
sylvain.chevallier@universite-paris-saclay.fr
raphael.camargo@ufabc.edu.br
marie-constance.corsi@inria.fr

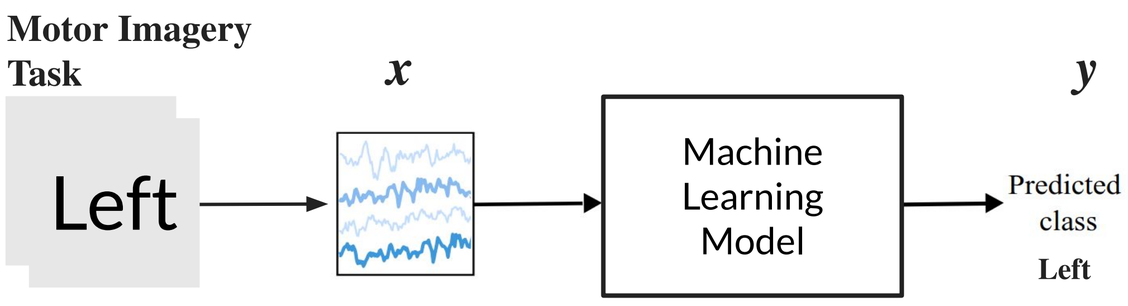
WHAT IS THE BEST MODEL FOR DECODING NEUROPHYSIOLOGICAL SIGNALS? DEPENDS ON HOW YOU EVALUATE

Bruno Aristimunha[1,2], Thomas Moreau[3], Sylvain Chevallier[1], Raphael Y. de Camargo[2], Marie-Constance Corsi[4]

1. Inria TAU, LISN-CNRS, Université Paris-Saclay,| 2. CMCC, UFABC, Santo André, Brazil | 3. Inria Mind, CEA, Université Paris-Saclay, France | 4. Sorbonne Université, ICM, CNRS, Inria NERV, Inserm, France

What is decoding neurophysiological signals?

Here, the neurophysiological time series ***x*** depends on the stimulus ***y***



When you train a machine learning model, you learn how to decode an task.

What is it benchmark?

Benchmarking is an ***emerging science***, and we understand it as the iron rule to tame anything goes. All disputes must be settled by competitive empirical testing:

1) Agree on metric;

2) Agree on benchmark data;

3) Compete (Compute).

Dataset and models (what we agree)

Here, we selected four motor-imagery datasets:

Dataset	Subjects	Channels	Sampling Rate (Hz)	Sessions	Tasks	Trials/Class	Epoch (s)
BNCI2014001	9	22	250	2	4	144	[2, 6]
BNCI2014004	9	3	250	5	2	360	[3, 7.5]
Weibo2014	10	60	200	1	7	80	[3, 7]
Zhou2016	4	14	250	3	3	160	[0, 5]

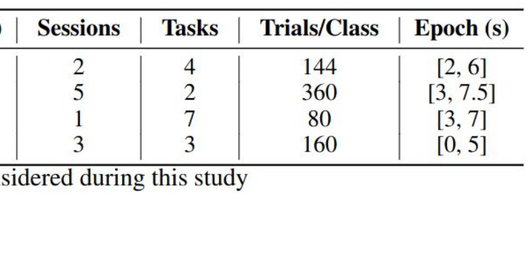
Table 1: Motor Imagery datasets considered during this study

And we selected 17 machine learning models:

NAME	YEAR	CATEGORY
BIOT	2023	DEEP
AUG-COV	2024	FC
ATCNet	2022	DEEP
AttentionNet	2024	DEEP
EEGITNet	2022	DEEP
EEGInception	2020	DEEP
EEGNetv4	2018	DEEP
ShallowFBCSPNet	2016	DEEP
TIDNet	2020	DEEP
Cov-CSP-LDA	2008	FC
Cov-CSP-LDA	2008	FC
Cov-FgMDM	2010	FC
Cov-MDM	2010	FC
Cov-Tang-LogReg	2010	FC
Fucone	2022	FC
Cov-Tang-SVM	2010	FC
LogVar-LDA	2008	FC
LogVar-SVM	2008	FC

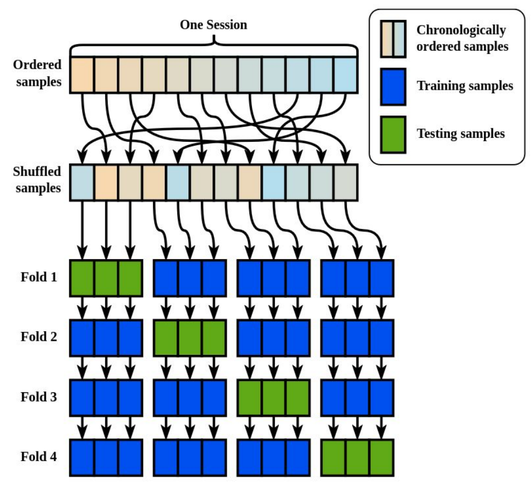
Deep learning

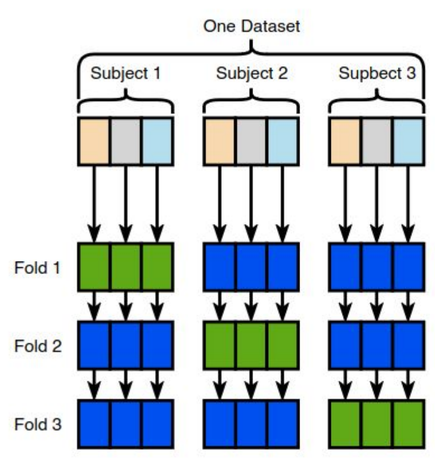
Functional Connectivity



How do we evaluate?

Inter vs Intra models



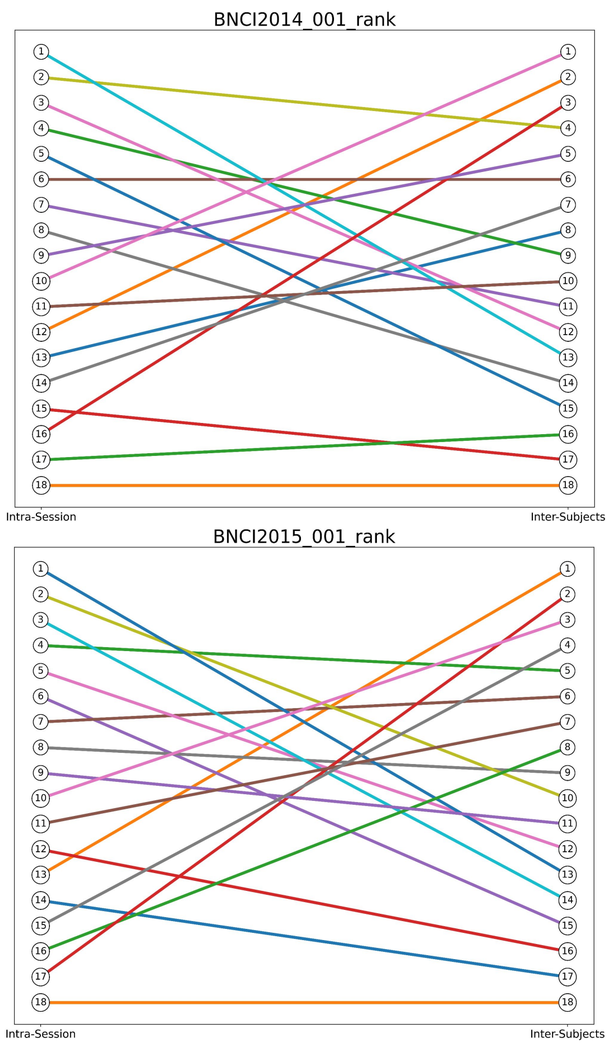


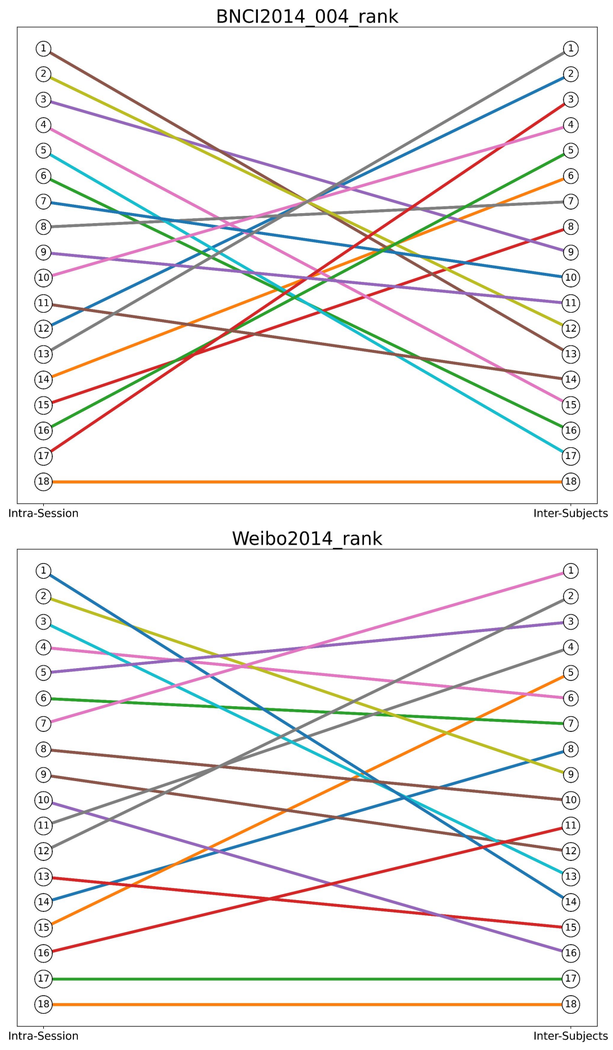
Experimental results

1. We can decode the time series!
   One subject for model | multi subject for model

solver_name	BNCI2014_001	BNCI2014_004	BNCI2015_001	Weibo2014	BNCI2014_001	BNCI2014_004	BNCI2015_001	Weibo2014
ATCNet	38.93±12.80	63.01±18.96	55.79±13.37	27.21±6.41	35.07±4.11	**74.78±4.46**	53.27±2.90	18.45±2.62
AttentionBaseNet	39.85±13.73	57.65±16.93	57.39±16.03	26.70±6.87	47.41±6.61	67.36±7.31	**66.03±5.57**	18.70±1.95
AUG-COV	67.47±18.67	70.22±18.04	81.43±14.08	46.53±10.92	33.70±3.77	60.95±6.27	59.12±5.17	18.46±4.11
BIOT	33.33±10.47	55.75±16.01	58.00±14.33	27.94±6.80	26.65±1.72	66.44±2.24	53.28±0.93	17.43±1.81
Cov-CSP-LDA_shr	61.53±17.93	71.15±17.55	78.46±15.14	36.83±11.66	32.95±5.68	65.24±2.88	54.42±2.06	16.67±3.23
Cov-CSP-LDA_svd	62.87±17.41	**72.29±16.20**	77.36±16.65	38.34±12.05	35.45±5.15	63.69±3.25	59.04±6.00	17.93±4.45
Cov-FgMDM	69.20±16.29	70.78±18.21	81.00±13.94	60.54±11.44	32.76±3.85	61.61±3.66	55.44±1.39	18.48±3.17
Cov-MDM	60.27±15.26	69.03±17.92	75.43±17.26	32.56±11.08	30.45±5.67	66.83±3.31	57.65±5.98	18.98±2.53
Cov-Tang-LogReg	70.19±16.49	72.10±17.45	81.96±14.49	64.36±9.41	36.35±6.75	64.56±3.31	57.00±2.69	18.30±2.91
Fucone	**70.27±15.49**	70.64±18.16	81.50±13.61	**63.25±10.96**	32.38±3.73	60.45±4.46	54.59±1.48	17.80±2.43
Cov-Tang-SVM	67.32±15.83	69.80±17.53	**82.36±13.89**	65.30±9.92	30.24±4.75	65.20±3.57	54.76±3.20	17.72±2.75
DUMMY	15.71±4.24	40.53±7.81	41.82±6.27	9.03±3.04	25.00±0.00	50.00±0.00	50.00±0.00	14.29±0.00
EEGInception	30.77±12.07	54.41±16.92	53.89±12.58	21.66±6.79	30.21±1.95	70.41±4.81	58.20±1.69	16.39±2.72
EEGNetv4	32.91±9.77	52.13±13.97	51.89±9.99	22.98±7.97	42.73±5.08	74.74±3.42	65.01±4.45	18.13±3.53
LogVar-LDA	56.36±14.77	69.02±17.51	72.43±16.84	51.55±8.31	35.69±5.47	64.80±5.49	55.98±3.30	20.29±1.84
LogVar-SVM	42.76±16.46	66.52±20.04	63.86±17.39	41.94±7.51	33.19±2.29	63.33±1.38	58.39±4.41	18.16±2.71
ShallowFBCSPNet	48.24±16.57	67.69±19.03	67.07±17.03	45.47±10.30	**56.09±6.17**	73.26±2.24	64.67±4.11	**27.07±4.91**
TIDNet	34.21±10.21	62.72±20.05	55.25±13.12	30.21±7.45	35.40±2.77	75.10±4.45	63.41±1.64	21.68±3.54

2. The model ranks change completely!





Take-Home insights

1) The way the evaluation ***is the devil in the details***!

2) The variance of ranks between datasets is large; the best model for one dataset isn't necessarily the best for all. New models should be evaluated across different datasets, and statistical conclusions should be drawn.

3) More data appears to improve deep learning models. It doesn't make sense to build deep learning models if we train one model per subject, traditional models will be better;

4) We need to optimize the models more, but we have a tradeoff in computational costs, just here in an exploratory study, we trained more than 10500 models.

Linkedin:

33rd Annual Computational Neuroscience Meeting*CNS 2024

b.aristimunha@gmail.com