# Across-subject ensemble-learning alleviates the need for large samples for fMRI decoding

Himanshu Aggarwal[1][0000−0003−2451−0470], Liza Al-Shikhley[1][0009−0002−0297−4231], and Bertrand Thirion[1][0000−0001−5018−7895]

Inria, CEA, Université Paris-Saclay, Palaiseau, 91120, France
himanshu.aggarwal@inria.fr
liza.alshikhley@gmail.com
bertrand.thirion@inria.fr
https://team.inria.fr/mind/

**Abstract.** Decoding cognitive states from functional magnetic resonance imaging is central to understanding the functional organization of the brain. Within-subject decoding avoids between-subject correspondence problems but requires large sample sizes to make accurate predictions; obtaining such large sample sizes is both challenging and expensive. Here, we investigate an ensemble approach to decoding that combines the classifiers trained on data from other subjects to decode cognitive states in a new subject. We compare it with the conventional decoding approach on five different datasets and cognitive tasks. We find that it outperforms the conventional approach by up to 20% in accuracy, especially for datasets with limited per-subject data. The ensemble approach is particularly advantageous when the classifier is trained in voxel space. Furthermore, a Multi-layer Perceptron turns out to be a good default choice as an ensemble method. These results show that the pre-training strategy reduces the need for large per-subject data.

**Keywords:** fMRI · Decoding · Ensemble-learning

## 1 Introduction

Brain *decoding* refers to the process of inferring cognitive states from an individual's brain signals. It is an important tool to understand how information processing is distributed in the brain [26] and to spot potential dysfunction in neurology or psychiatry [19]. It entails acquiring brain signals using neuroimaging techniques such as functional magnetic resonance imaging (fMRI) and then training a classifier to learn the mapping from brain signals to cognitive states. However, given that acquiring such data is expensive and time-consuming, the number of features (voxels) is usually much larger than the number of samples (repetitions of stimulus presentation). This high feature-to-sample ratio, together with low data signal-to-noise ratio is detrimental to the quality of predictions [11], hence solutions to increase the amount of available data are needed.

*Across-subject correspondence issues:* One way to increase the sample size could simply be to collect data from many individuals and then train classifiers across these individuals to decode the brain activity of a new individual. However, it has been shown that spatio-temporal patterns of brain activity generalize poorly across individuals [8] leading to sub-optimal decoding accuracy [20]. Similar issues also exist in decoding brain signals from other neuroimaging techniques like magneto-/electro-encephalography (M/EEG) that have direct applications in brain-computer interfaces (BCI) [23]. A common solution in M/EEG-BCI research is aimed towards efficiently using the available information *within-subjects* by combining classifiers via voting or stacking [15,14]. The combination of classifiers can be done at the feature level or the decision level [14].

*Ensemble-learning:* One recent study [9] showed that fMRI-based brain activity *encoding* benefits from a linear combination of other subjects' predicted response vectors. However, to our knowledge, this approach of combining models and learning from an *ensemble* of individuals has not been investigated within the context of decoding, which is addressed in the present study.

## 2    Methods

*Ensemble-learning (by stacking):* In a conventional decoding setting, a classifier is trained to learn the mapping between stimuli labels and a given feature space. Let $\mathbf{X}_1, \cdots, \mathbf{X}_N \in \mathcal{X}^N$ be the fMRI datasets acquired in $N$ subjects, together with the corresponding labels $\mathbf{y}_1, \cdots, \mathbf{y}_N \in \mathcal{Y}^N$. Here, we present an ensemble approach where we first *(i)* pre-train separate classifiers $f_i$ for each subject $i \in [\![N-1]\!]$ to learn the mapping between their respective feature space $\mathcal{X}$ and stimuli labels in $\mathcal{Y}$

$$\forall i \in [\![N-1]\!], f_i : \mathbf{X}_i \longrightarrow \mathbf{y}_i, \tag{1}$$

and then *(ii)* train another classifier $g$ to learn the mapping between the predicted labels $f_i(\mathbf{X}_N)$ from the pre-trained classifiers $f_i$, and the true labels $\mathbf{y}_N$ for the remaining $N^{th}$ subject's features $\mathbf{X}_N$ (Fig. 1).

$$g : (f_1(\mathbf{X}_N), f_2(\mathbf{X}_N), ..., f_{N-1}(\mathbf{X}_N)) \longrightarrow \mathbf{y}_N \tag{2}$$

We do this in each subject and report the average cross-validated accuracy.

*Theory: different bias-variance decompositions.* The interest of this ensembling can be understood by considering bias-variance decompositions. For simplicity, we consider Ridge regression and rely on known bounds [12]. Dataset specific models $(f_i)_{i \in [\![N]\!]}$ yield a squared bias $\propto \sum_{j=1}^{N_{samples}} \frac{\lambda_j}{(1+\lambda_j/\lambda)^2}(f_i^j)^2$, where $f_i^j$ is the reponse magnitude to the $j^{th}$ dimension of the input, $(\lambda_j)$ are the eigenvalues of $\mathbf{X}_i$'s covariance matrix and $\lambda$ is the regularization parameter; the prediction variance is $\propto \frac{var(\mathbf{y})d(\lambda)}{N_{samples}}$, where $d(\lambda) = \sum_{j=1}^{N_{samples}} \left(\frac{\lambda_j}{\lambda_j+\lambda}\right)^2$ is the effective dimension of the Ridge model. For optimal $\lambda$, both terms are balanced, hence the
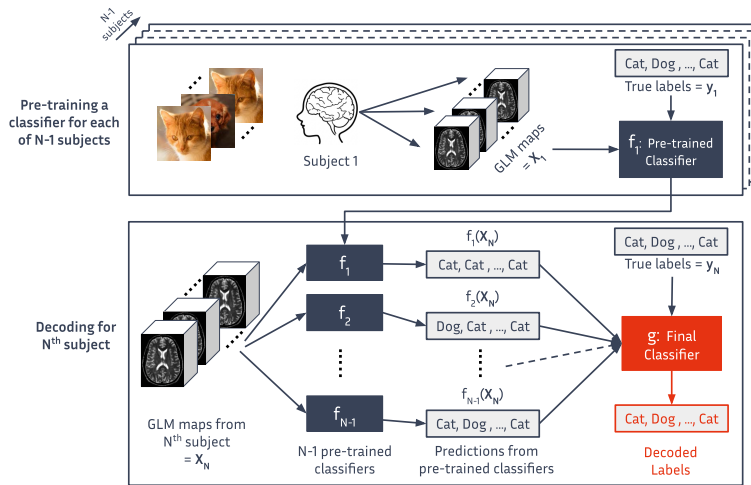
**Fig. 1. Ensemble-learning (by stacking):** Ensemble-learning (by stacking) involves (top) pre-training separate classifiers for each of $N-1$ subjects and then (bottom) training a final classifier to learn the mapping between predictions from each pre-trained classifier and the true labels for $N^{th}$ subject.

squared prediction error is $\propto \frac{var(\mathbf{y})d(\lambda)}{N_{samples}}$. If the ensembling model $g$ is ordinary least squares, it has no estimation bias, but a model bias $\propto \frac{var(\mathbf{y})}{N}$ due to the projection of the feature space on the $(f_i(\mathbf{X}_N))_{i \in [\![N-1]\!]}$ span, and a variance $\propto \frac{var(\mathbf{y})N}{N_{samples}}$. This leads to a squared error of order $\frac{var(\mathbf{y})}{N} + \frac{var(\mathbf{y})N}{N_{samples}}$.

As a consequence, we can outline three regimes: *i)* if $N$ is very small ($\mathcal{O}(1)$), then the ensemble model is affected by large modeling error $\frac{var(\mathbf{y})}{N}$ and is not competitive; *ii)* if $N$ is larger, e.g. $\mathcal{O}(\sqrt{N_{samples}})$ then both error terms are balanced, making the ensemble model competitive with respect to the standard model ($N < d(\lambda)$) ; *iii)* when $N$ becomes as large as $N_{samples}$, the ensemble model is dominated by variance; it thus requires regularization and may no longer be competitive. Overall, we can expect accuracy gains when $1 \ll N \ll N_{samples}$.

*Datasets and tasks:* We compare the ensemble approach against the conventional one in five different fMRI datasets and tasks with different characteristics (Table 1). Each of these datasets has 4 to 6 different conditions (or classes) of stimuli to be decoded. All these datasets are publicly available and their pre-processing pipelines are described in detail in their respective publication. We further use Nilearn [1] to extract brain signals, detrend and standardize these preprocessed data. These datasets have varying voxel resolution, thus we down-sampled them to $3mm$ isotropic resolution for consistency, and to reduce computational costs. Additionally, we applied a $5mm$ Gaussian smoothing kernel to improve the signal-to-noise ratio. Finally, we fit a general linear model (GLM) to each dataset to derive trial-by-trial effect size maps using the Least Squares

**Table 1. List of fMRI datasets and corresponding tasks used in the study:** $N_{samples}$ refers to the number of samples per subject, $N_{subjects}$ to the number of subjects in the cohort and $N_{classes}$ to the number of classes in the prediction task.

| Dataset | Task | $N_{samples}$ | $N_{subjects}$ | $N_{classes}$ | Stimuli labels |
|---|---|---|---|---|---|
| Neuromod [2] | Visual n-back | 50 | 4 | 4 | images of body/face/place/tools |
| AOMIC [21] | Emotion anticipation [16] | 61 | 203 | 4 | negative/neutral emotion images and cue for negative/neutral emotion images |
| Forrest [10] | Music genre perception [3] | 175 | 10 | 5 | ambient/country/metal/rocknroll/symphonic music |
| BOLD5000 [4] | Image-Net [6] image viewing | 332 | 3 | 4 | images of furniture/vehicle/animal/person |
| RSVP-IBC [18] | RSVP language [13] | 360 | 13 | 6 | type of text: jabberwocky/complex/simple/word list/pseudoword list/consonant strings |

Separate approach [25]. These trial-by-trial GLM effect size maps are used as features for decoding.

*Decoding settings:* In this work, we compare the two decoding approaches in two different feature spaces: the image voxel space (50k voxels) and a low-resolution space, based on the 1024-dimensional DiFuMo decomposition [5]. These DiFuMo features are obtained by regression using Nilearn [1]. Furthermore, we also compare three different classifiers for decoding: Multi-layer Perceptron (MLP), linear Support Vector Classifier (SVC), and Random Forest [17]. We use the default Scikit-learn parameters for each classifier with a few modifications. For MLP, we use 100 hidden layers, ReLU activation, ADAM solver and 1000 iterations. For SVC, we use $l2$ penalty, and square hinge loss, and the algorithm to solve dual optimization is set to automatic. For Random Forest we use 500 trees, and for the maximum depth of a tree, the nodes are expanded until all leaves are pure or until all leaves contain less than 2 samples. Note that only the final classifier ($g$ in Eq. 2) is switched between the three model families, and the pre-trained classifiers ($f_i$ in Eq. 1) are always linear SVC with $l2$ penalization as this remains the best-performing classifier for this type of data. Results with $l1$ penalization for $f_i$ are given in Supplementary Fig. 2.

*Varying the training-set size:* Within each dataset, we keep 90% of data for training and 10% for testing. We vary the size of the training set over 10 geometrically increasing sub-samples of that initial 90% training split and always test the trained model on the same 10% testing split. We do this for 20 different cross-validation train-test splits. Note that in the ensemble approach, while pre-training the classifiers, we use all the samples available in each subject.

*Varying the number of subjects in the ensemble:* Next, we randomly sample a subset of subjects from each dataset and only use the pre-trained classifiers from these subjects to train the final classifier. For each subset of subjects, we also vary the training set size as described in the previous paragraph. We do this for 5 different cross-validation train-test splits, such that each split has a different subset of subjects, whenever possible.

*Evaluation metric:* We use the average (balanced) accuracy across all subjects in the dataset and twenty cross-validation splits, over ten training set sizes. Balancing is done to account for the class imbalance for BOLD5000, where the metric used is the average recall across classes. To get an estimate of the uncertainty in the average accuracy, we use a bootstrap approach as implemented in the Python library Seaborn. The error bars represent a 95% interval of the bootstrap distribution.

*Code availability:* All the code used in this study is available at:
`https://github.com/man-shu/ensemble-fmri`

## 3 Results

*The Ensemble approach outperforms conventional decoding:* Fig. 2 shows the average decoding accuracies for the conventional and ensemble approaches across all datasets and decoding settings. The ensemble approach outperforms the conventional approach in all datasets and decoding settings, except BOLD5000. The best average performance gain is observed in the Neuromod dataset with 20% gains between best-performing conventional and ensemble approaches (Ensemble, Voxels, MLP vs. Conventional, DiFuMo, LinearSVC). Using full-voxel feature space for pre-training is more beneficial than a reduced feature space, in all datasets, irrespective of the classifier used. Comparisons between classifiers, however, are not straightforward. In general, we see that MLP is the best-performing classifier — at least it performs similarly to the best-performing classifier in all datasets and decoding settings. However, in datasets where 10 or more subjects are available, and the number of samples is higher (as in Forrest and RSVP-IBC), Random Forests outperform linear SVC. On the other hand, in datasets where the number of subjects is higher (as in AOMIC) or when data is scarce (as in Neuromod), linear SVC outperforms Random Forest.

In addition to the gain in accuracy, we could also extract feature importance scores for each dataset/task (Supplementary Fig. 1). These scores help explore the spatial patterns of brain activity that are most informative for decoding and could hence be informative for understanding the cognitive processes. For example, for the visual n-back task of Neuromod, the most informative voxels are located in the visual cortex; for the music genre perception task of Forrest, in the auditory cortex; and for the RSVP language task of RSVP-IBC, in the language processing areas.

*The case of scarce data and many subjects:* We further investigate the effect of the number of training samples per class on accuracy. The *gain in accuracy*
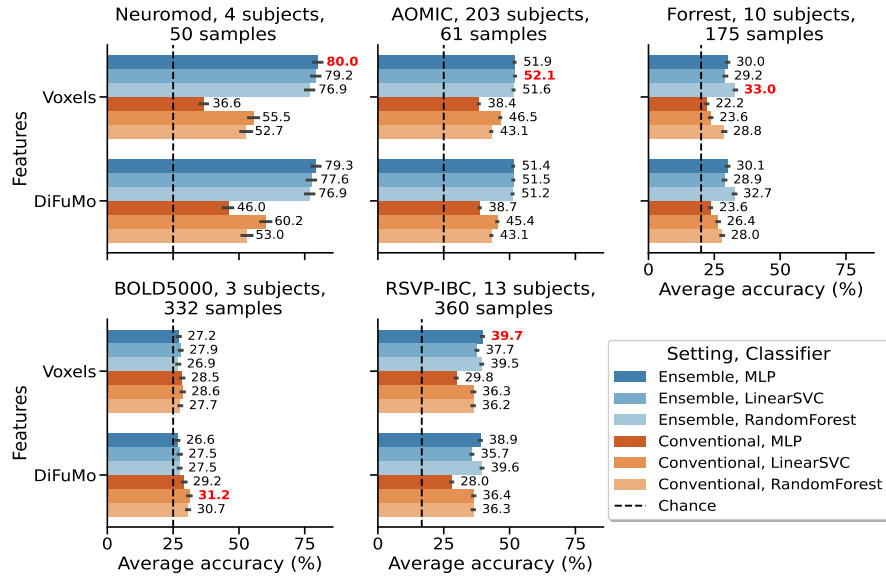
**Fig. 2. Average decoding accuracy:** Each plot represents a different dataset (along columns). The average decoding accuracy is plotted along the x-axis. The averages are across all training sizes, subjects and 20 cross-validation splits. The error bars represent a 95% confidence interval of the bootstrap distribution. The horizontal line represents the chance level of accuracy.

displayed in Fig. 3 is the difference in accuracy between the ensemble and conventional approaches. In all datasets, with MLP as the final classifier, the gains are much higher relative to the other two classifiers. This is due to the poor performance of MLP in the conventional approach in the scarce data regime. On the other hand, with linear SVC and Random Forest, gains are observed when up to 10 samples per class are available for training in all datasets, except BOLD5000 (Fig. 3). When more than 10 samples per class are available, the ensemble approach performs similarly to the conventional approach in all datasets, except BOLD5000 and RSVP-IBC. Note that both datasets have the largest number of samples per class and this observation indicates that the ensemble approach is only beneficial when the number of samples per class is low. Furthermore, it should be noted that the BOLD5000 dataset has the lowest number of subjects (3) among all datasets and the loss of performance in this case shows that the availability of very few subjects hampers the ensemble approach even when the data is scarce. Therefore, we also examined the effect of the number of subjects on the gain in accuracy.

Fig. 4 shows that gains in accuracy increase with an increasing number of subjects in the ensemble in all datasets and decoding settings. Again, the gains are higher with MLP as the final classifier and are positive even when there is only
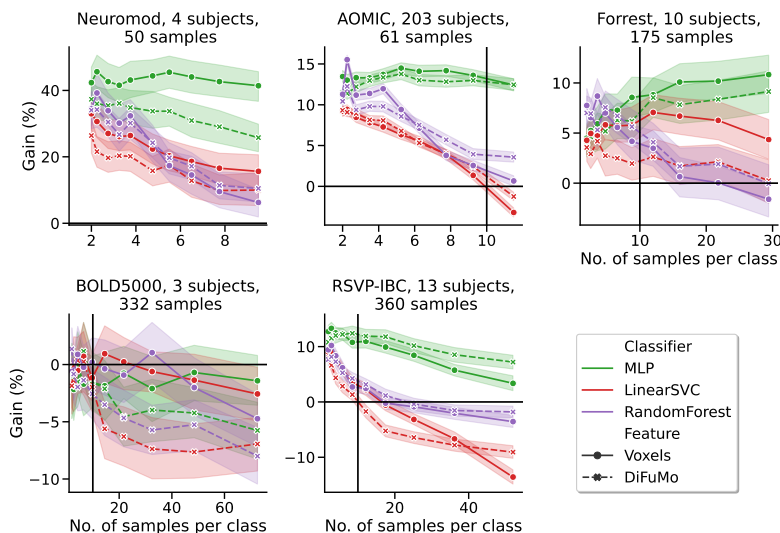
**Fig. 3. Gain in decoding accuracy when varying the number of training samples per class:** Each plot represents a different dataset (along columns). The y-axis shows the average percent gain in decoding accuracy (accuracy of ensemble - accuracy of conventional) across all subjects and 20 cross-validation splits. On the x-axis, training size is reported as the number of samples per class in each cross-validation split. The confidence intervals represent 95% confidence interval of bootstrap distribution. The horizontal line represents no average gain in accuracy and the vertical line, 10 samples per class.

one subject in the ensemble. For the other two classifiers, the AOMIC dataset with the largest number of subjects (203) shows that the gains become positive starting from $3-4$ subjects and saturate at around 10 subjects. The saturation is observed even with the MLP classifier. The RSVP-IBC dataset with the largest number of samples (360) and stimulus classes (6) shows that when more samples are available and the classification task is more complex, the gains continue to increase even with more than 10 subjects in the ensemble. Additionally, we tested the ensemble approach with linear SVC with $l1$ penalization during pretraining. The average decoding accuracies are lower relative to $l2$ penalization, in all datasets, except AOMIC and RSVP-IBC (Supplementary Fig. 2) – the two datasets with the largest number of subjects. This indicates that the $l1$ penalization is beneficial when the number of subjects is large.

## 4    Discussion

In this work, we set out to investigate the potential of an ensemble approach towards decoding cognitive states from fMRI signals. This approach leverages patterns learned from an ensemble of other subjects. We found that it outper-
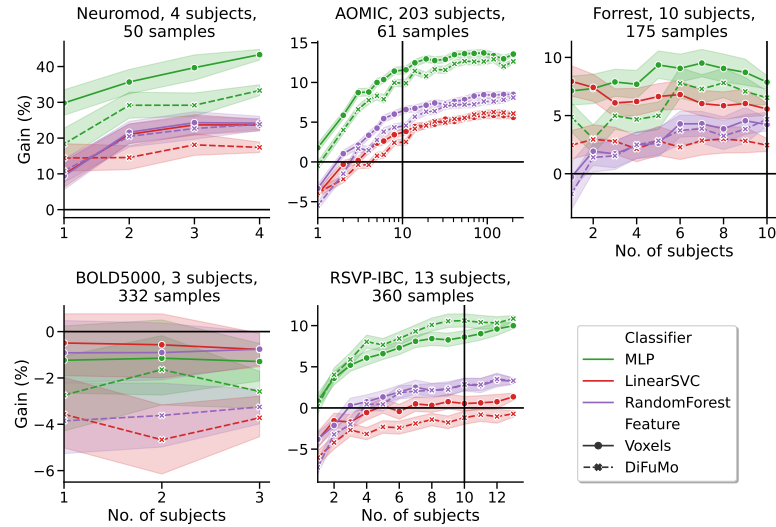
**Fig. 4. Gain in decoding accuracy over a varying number of subjects in the ensemble:** Each plot represents a different dataset (along columns). The x-axis represents the number of subjects used in the ensemble method. The y-axis represents the average percent gain in decoding accuracy (accuracy of ensemble - accuracy of conventional) across all training sizes and 5 cross-validation splits. The confidence intervals represent 95% interval of bootstrap distribution. The horizontal line represents no average gain in accuracy and the vertical line at 10 subjects in the ensemble.

forms the conventional decoding approach in four of the five different datasets considered. Particularly, as predicted by bias/variance analysis, this approach leads to gains in accuracy when data is scarce, and the gains increase with an increasing number of subjects in the ensemble. The only dataset where the ensemble approach did not outperform the conventional approach was BOLD5000, which has the lowest number of subjects (3) among all datasets. On average, using the full-voxel space as features is more beneficial than a reduced feature space, in all datasets, irrespective of the classifier used. Using MLP as the final classifier is a safe choice, as it is the best-performing classifier or at least performs similarly to the best-performing classifier in all datasets and decoding settings. The relative performance of linear SVC and Random Forest is sensitive to the number of subjects and samples available for training.

Given that decoding is a central tool in cognitive mapping and clinical applications, the need for large sample sizes is a major bottleneck. Therefore, the present study serves as a proof of concept that combining pre-trained classifiers from different individuals allows for efficient use of available information and alleviates the need for large sample sizes in decoding fMRI brain signals. On average, we observe that due to the gain in accuracy with this approach, the number of samples required for training can be reduced by $5 - 10$ samples per

class, depending on the dataset. Moreover, given that this approach provides a gain in prediction accuracy in scarce data conditions, it can also be beneficial in real-time BCI applications [7,22] where initial data scarcity is a major issue. Furthermore, this method could be particularly useful in basic cognitive research that uses task fMRI, where a given cognitive domain could not be decoded accurately due to a small sample size or low signal-to-noise ratio in conventional settings.

The primary limitation of the approach is that it cannot work when very few subjects are available – as seen in the case of the BOLD5000 dataset. However, there is some evidence that similar approaches are immune to cross-dataset covariate shifts [9]. Therefore, by combining classifiers pre-trained on a large set of subjects as presented in this work, one can leverage large-scale public datasets like Amsterdam Open MRI Collection (AOMIC) [21]. The second limitation is that the gains with ensembling are only obtained in the case of scarce per-subject data. In the present study, except for trying $l1$ and $l2$ penalizations for linear SVC classifiers during pre-training, we did not explore any other pre-training strategies. Pre-training deep learning models instead could be beneficial when larger sample sizes are available. Finally, it would also be interesting to investigate if the approach can benefit even further from functional alignment techniques [24] that improve across-subject correspondence.

**Disclosure of Interests.** The authors declare no competing interests.

# References

1. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G.: Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics **8** (2014). `https://doi.org/10.3389/fninf.2014.00014`
2. Bellec, P., Boyle, J.: Bridging the gap between perception and action: The case for neuroimaging, AI and video games (Jan 2024). `https://doi.org/10.31234/osf.io/3epws`
3. Casey, M., Thompson, J., Kang, O., Raizada, R., Wheatley, T.: Population Codes Representing Musical Timbre for High-Level fMRI Categorization of Music Genres. In: Langs, G., Rish, I., Grosse-Wentrup, M., Murphy, B. (eds.) Machine Learning and Interpretation in Neuroimaging. pp. 34–41. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2012). `https://doi.org/10.1007/978-3-642-34713-9_5`
4. Chang, N., Pyles, J.A., Marcus, A., Gupta, A., Tarr, M.J., Aminoff, E.M.: BOLD5000, a public fMRI dataset while viewing 5000 visual images. Scientific Data **6**(1), 49 (May 2019). `https://doi.org/10.1038/s41597-019-0052-3`
5. Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K.J., Wassermann, D., Thirion, B., Mensch, A.: Fine-grain atlases of functional modes for fMRI analysis. NeuroImage **221**, 117126 (Nov 2020). `https://doi.org/10.1016/j.neuroimage.2020.117126`
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (Jun 2009). `https://doi.org/10.1109/CVPR.2009.5206848`
7. Fede, S.J., Dean, S.F., Manuweera, T., Momenan, R.: A Guide to Literature Informed Decisions in the Design of Real Time fMRI Neurofeedback Studies: A Systematic Review. Frontiers in Human Neuroscience **14** (2020). `https://doi.org/10.3389/fnhum.2020.00060`
8. Gratton, C., Laumann, T.O., Nielsen, A.N., Greene, D.J., Gordon, E.M., Gilmore, A.W., Nelson, S.M., Coalson, R.S., Snyder, A.Z., Schlaggar, B.L., Dosenbach, N.U.F., Petersen, S.E.: Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation. Neuron **98**(2), 439–452.e5 (Apr 2018). `https://doi.org/10.1016/j.neuron.2018.03.035`
9. Gu, Z., Jamison, K., Sabuncu, M., Kuceyeski, A.: Personalized visual encoding model construction with small data. Communications Biology **5**(1), 1–12 (Dec 2022). `https://doi.org/10.1038/s42003-022-04347-z`
10. Hanke, M., Dinga, R., Häusler, C., Guntupalli, J.S., Casey, M., Kaule, F.R., Stadler, J.: High-resolution 7-Tesla fMRI data on the perception of musical genres – an extension to the studyforrest dataset. F1000Research **4**, 174 (Jun 2015). `https://doi.org/10.12688/f1000research.6679.1`
11. Haynes, J.D.: A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. Neuron **87**(2), 257–270 (Jul 2015). `https://doi.org/10.1016/j.neuron.2015.05.025`
12. Hsu, D., Kakade, S.M., Zhang, T.: Random design analysis of ridge regression. Foundations of Computational Mathematics **14**(3), 569–600 (2014)
13. Humphries, C., Binder, J.R., Medler, D.A., Liebenthal, E.: Syntactic and Semantic Modulation of Neural Activity during Auditory Sentence Comprehension. Journal of Cognitive Neuroscience **18**(4), 665–679 (Apr 2006). `https://doi.org/10.1162/jocn.2006.18.4.665`

14. Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., Yger, F.: A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update. Journal of Neural Engineering **15**(3), 031005 (Jun 2018). `https://doi.org/10.1088/1741-2552/aab2f2`
15. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain–computer interfaces. Journal of Neural Engineering **4**(2), R1–R13 (Jun 2007). `https://doi.org/10.1088/1741-2560/4/2/R01`
16. Oosterwijk, S.: Choosing the negative: A behavioral demonstration of morbid curiosity. PLOS ONE **12**(7), e0178399 (Jul 2017). `https://doi.org/10.1371/journal.pone.0178399`
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.: Scikit-learn: Machine Learning in Python. the Journal of machine Learning research **12**, 2825–2830 (2011)
18. Pinho, A.L., Amadon, A., Gauthier, B., Clairis, N., Knops, A., Genon, S., Dohmatob, E., Torre, J.J., Ginisty, C., Becuwe-Desmidt, S., Roger, S., Lecomte, Y., Berland, V., Laurier, L., Joly-Testault, V., Médiouni-Cloarec, G., Doublé, C., Martins, B., Salmon, E., Piazza, M., Melcher, D., Pessiglione, M., van Wassenhove, V., Eger, E., Varoquaux, G., Dehaene, S., Hertz-Pannier, L., Thirion, B.: Individual Brain Charting dataset extension, second release of high-resolution fMRI data for cognitive mapping. Scientific Data **7**(1), 353 (Oct 2020). `https://doi.org/10.1038/s41597-020-00670-4`
19. Sarraf, S., DeSouza, D.D., Anderson, J., Tofighi, G., Initiativ, f.t.A.D.N.: DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI (Jan 2017). `https://doi.org/10.1101/070441`
20. Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M., Just, M.A.: Using fMRI Brain Activation to Identify Cognitive States Associated with Perception of Tools and Dwellings. PLOS ONE **3**(1), e1394 (Jan 2008). `https://doi.org/10.1371/journal.pone.0001394`
21. Snoek, L., van der Miesen, M.M., Beemsterboer, T., van der Leij, A., Eigenhuis, A., Steven Scholte, H.: The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. Scientific Data **8**(1), 85 (Mar 2021). `https://doi.org/10.1038/s41597-021-00870-6`
22. Sorger, B., Goebel, R.: Real-time fMRI for brain-computer interfacing. In: Handbook of Clinical Neurology, vol. 168, pp. 289–302. Elsevier (2020). `https://doi.org/10.1016/B978-0-444-63934-9.00021-4`
23. Thompson, M.C.: Critiquing the Concept of BCI Illiteracy. Science and Engineering Ethics **25**(4), 1217–1233 (Aug 2019). `https://doi.org/10.1007/s11948-018-0061-1`
24. Thual, A., Tran, Q.H., Zemskova, T., Courty, N., Flamary, R., Dehaene, S., Thirion, B.: Aligning individual brains with fused unbalanced Gromov Wasserstein. Advances in Neural Information Processing Systems **35**, 21792–21804 (Dec 2022)
25. Turner, B.O., Mumford, J.A., Poldrack, R.A., Ashby, F.G.: Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. NeuroImage **62**(3), 1429–1438 (Sep 2012). `https://doi.org/10.1016/j.neuroimage.2012.05.057`
26. Zhang, Y., Tetrel, L., Thirion, B., Bellec, P.: Functional annotation of human cognitive states using deep graph convolution. NeuroImage **231**, 117847 (May 2021). `https://doi.org/10.1016/j.neuroimage.2021.117847`