Self-Supervised Learning for Functional Brain Networks identification in fMRI from Healthy to Unhealthy Patients

Lukman Ismaila, Pejman Rasti, Jean-Michel Lemée, David Rousseau

Self-Supervised Learning for Functional Brain Networks identification in fMRI from Healthy to Unhealthy Patients

Lukman E. Ismaila
LARIS, UMR INRAe, IRHS
Université d'Angers
Angers, France

Pejman Rasti
CERADE, LARIS
ESAIP, Université d'Angers
Angers, France

Jean-Michel Lemée
Service de Neurochirurgie
CHU d'Angers
Angers, France

David Rousseau*
LARIS, UMR INRAe, IRHS
Université d'Angers
Angers, France

***Abstract*—Resting State Functional Magnetic Resonance Imaging (rs-fMRI) technique is gaining more attention among medical practitioners because, it allows recognition of functional brain networks and is very suitable for complex situations where the participation of the patients is not required. This approach is also interesting for non-invasive medical imaging where healthy subjects can be enrolled very easily during the data acquisition process. However, one of its limitations is that the clinicians must manually annotate the image data. While no clinical use of this annotation is needed at any stage of neurosurgical procedure, this process is often time consuming and can only be carried our by domain experts. We investigate the possibility to perform self-supervision from healthy subject data without the need of image annotation, followed by transfer learning from the models trained on some pretext task. The result of self-supervision is shown to bring about 3% increase in performance without the effort and time of manual annotation of fMRI data by expert.**

***Index Terms*—Self-supervision, image classification, medical imaging, functional brain network, fMRI, transfer learning**

I. Introduction

Supervised machine learning experiences a large success in computer vision driven medical imaging nowadays [1]. However, there are some well-known limitations in the application of these data driven methods. One of these limitations is the usual lack of large annotated data sets which may not be available because they correspond to rare disease, or because the international community maintain limited distribution of public dataset, or because human expertise for the annotation of the dataset is limited.

There are several workarounds to compensate for the limited availability of dataset [2], [3]. These include few-shot learning, creation of artificial data, generative models, or data enhancement. Transfer learning, another common strategy, makes use of models that have already been trained on comparable dataset. Very recently [4], we showed the possibility to use such transfer learning from healthy subjects to unhealthy patients. This is very interesting indeed for medical imaging modalities which are purely non-invasive and therefore for which it is rather easy to enroll healthy control. This was illustrated in [4] for a task of functional brain network identification in resting state functional magnetic resonance imaging (fMRI). A significant gain in classification performance was obtained because the brain tumour of the unhealthy patients was found to have very limited impact on the resting state fMRI (rs-fMRI) signals.

We propose a follow up of the recent study of [4] in this communication. A limitation in the transfer learning from healthy patients to unhealthy patient is the need of manually annotation of the healthy patients. This annotation is time consuming while it has to be performed on patients for which there is clearly no clinical interest. To avoid this unnecessary step while trying to take benefit from the similarity between healthy subjects and unhealthy patients by transfer learning, we propose to investigate the possibility of self-supervision for the task targeted in [4].

Self-supervision, is a machine learning method which learns from unlabeled sample data [5]. It can be regarded as an intermediate form between supervised and unsupervised learning. It is usually based on an artificial neural networks. The training of the network is performed in two stages. First, a pretext task is solved based on pseudo-labels which contributes to initialize the network weights. Secondly, the target task is performed with supervised learning but with much fewer need of annotation due to the initialisation from the weights trained on the pretext task. Self-supervision is now applied in all fields of computer vision but recently, began to receive consideration for fMRI related data [6], [7]. In [6] a regression task to predict the fatigue from patient based on their fMRI patterns is targeted. In [7] images are generated from fMRI patterns after visualisation of the images by healthy patients. In this work, we explore the value of self-supervision from healthy to unhealthy data in the use case of [4] where functional brain networks have to be identified.

II. Materials and Methods

Database

This study is a single-center prospective, open-label trial that adheres to regulations and ethical standards for clinical research and has been approved by the local ethics committee (Comité de protection des personnes Ouest II, decision reference CPP 2012-25). We collected data from 55 unhealthy

patients and 81 healthy people. The data from healthy participants were collected from regular volunteers, whereas the data from unhealthy patients were collected from people with brain tumors. The given binary lesion mask shows where the lesion is. In [8], a thorough overview of the unhealthy population is given. Eighty-one healthy participants, ranging in age from 23 to 38, including 36 women and 45 men, completed written informed consent forms. Fifty-five patients with brain lesions received preoperative fMRI language mapping and perioperative cortical mapping of brain regions involved in eloquent brain language under awake settings at the University Hospital of Angers, the Department of Neurosurgery (CHU Angers). Before enrolling participants in this study, participant permission was acquired.

For both healthy and unhealthy data, we retrieved 55 features from independent component analysis (ICA) with a focus on 7 brain features. Determining the total number of components (TNC) to use in ICA in resting-state fMRI is one of the main challenges, which can result in suboptimal decompositions with the fusion of multiple networks in low TNC cases or, in high TNC situations, the division of a functional network into several components. [9], [10]. Based on earlier research, we examined 55 ICs across all patients to establish functional brain networks. [11], [12].

The primary Intrinsic Connectivity Network (ICN) found and reported in resting-state fMRI literature is represented by the seven chosen brain characteristics. These brain features correspond to 7 biological networks of the brain, which are the Salience Network (SAL), Language Network (LANG), Default Mode Network (DMN), Ventral Attention Network (VAN), Right Fronto-parietal Control Network (rFPCN), Left Fronto-parietal Control (lFPCN), Dorsal Attention Network (DAN). These particular networks were chosen for the DMN based on their inter-individual variability, which makes them difficult to find using detection techniques, to serve as a control for the others. These networks match links in well-known cognitive networks that have been used to the design of preoperative procedures. [10], [13]. The connection networks between rs-fMRI and different fMRI data collecting and processing approaches were also shown to be consistent [14]. In the algorithm training and automatic identification processes, functional networks with fixed locations, such as the visual cortex, sensory, or motor were not taken into account. Domain experts assigned labels to each healthy and unhealthy data file, and these labels were utilized to categorize each image into the appropriate network class. In addition to the two network picture variations offered for both healthy and unhealthy data, as shown in figure 2, unhealthy data also contains information about the brain tumor, as seen in figure 1 and explained in table I.

TABLE I: Description of Unhealthy patient database

	Description of the Image Data for the Unhealthy Patient	
	Files	**Description**
1	Lesion.nii	This file contains the binary mask for each patient's specific brain tumor.
2	Grey Matter mask (mrwp1)	It is the grey matter mask (helpful since all activation occurs in the grey matter).)
3	White Matter mask (mrwp2)	It serves as the white matter mask (no activation inside the white matter, but it may be a useful method for estimating the deformations of the brain caused by the tumor and the peritumor edema.)
4	Cerebrospinal fluid mask (mrwp3)	The cerebrospinal fluid mask (similar to white matter, no activation within, but perhaps useful to measure brain deformations))
5	Whole brain-white grey matter (wms)	White and gray matter across the whole brain in a T1 anatomical MRI sequence with the patient's skin and skull clipped
6	Whole brain (wmrs)	This is a depiction of the complete cerebrospinal fluid of the brain, together with the skull and skin.

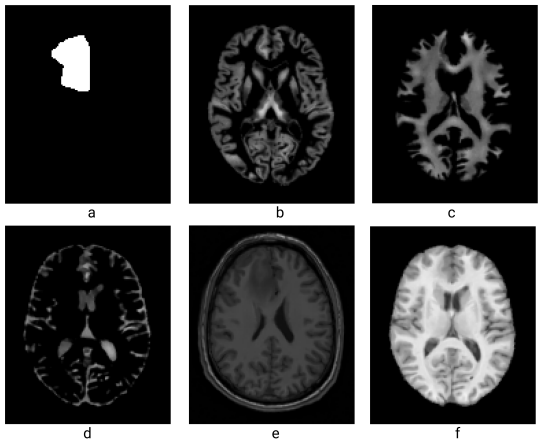
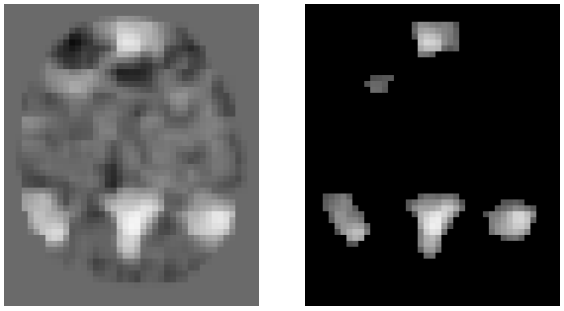


Fig. 1: (Visualization of components from unhealthy data: **a** represents the lesion, **b** represents the grey matter, **c** represents the white matter, **d** represents the cerebrospinal fluid, **e** represents the whole brain (white and grey matter), and **f** represents the cerebrospinal fluid alone).

Data acquisitions and preprocessing

All fMRI data acquisitions were performed using a 3 Tesla MRI (Siemens Medical Systems, Magnetom Skyra, and Erlangen) with a slice thickness of 4mm each, resulting in voxel sizes of $3 \times 3 \times 4\ mm^3$ and consequently a multichannel 3-D image of $42px \times 51px \times 34channels$. For each subject, the following fMRI sequences were acquired: one anatomical 3D T1, one resting-state acquisition, and two task-induced activity. At the time of the fMRI acquisition and throughout the surgery, neither any of the patients nor the healthy volunteers who were included had any linguistic impediment. In order to allow for the auto-adjustment of the magnetic field gradients, the first three volumes recorded in each series were dropped. The Anatomy, SPM8, and VBM 12 toolboxes of MatLab (The MathWorks, Natick, MA) were used for data preparation. Realignment to the first volume of the first session, slice-timing correction, and unwrapping to correct head motions and magnetic distortions were the processes used in the preparation of the fMRI data. After segmenting the images, the template from the Montreal Neurological Institute [15] was used to normalize them. Each patient's rs-fMRI data was segmented into 55 spatial independent components (ICs) using an intrinsic connection network spatial independent component analysis (SICA) technique that employed a modified version of the infomax algorithm running in Matlab. [16], [17]. ICs are 3D fMRI activation volumes of brain regions that exhibit spontaneous synchronized activity. Without any dispute, the reference fMRI identification of brain networks was completed

manually for each participant by two experienced and independent reviewers. We chose seven key networks of LANG, DMN, SAL, VAN, lFPCN, DAN, and rFPCN from the 55 produced ICs for each patient based on fMRI spatial distribution and activation peaks of these activations. There were two versions of the annotated photos: complete gray level images (connectivity map) and equivalent thresholded image copy. Figure 2 illustrates a DMN network picture sample. At the cluster level, individual spatial components were thresholded at z = 2, corresponding to the 5% most active voxels in each intrinsic connection network. This approach is consistent with the literature and enables for the identification of the anatomical location of active brain regions despite background activation noise. [18].



(a) gray level image (b) Thresholded image

Fig. 2: Visualization of image data variants with Example from Default Mode Network (DMN)

Identification of functional networks through machine learning algorithms

Fewer than half of the 55 ICs discovered using the SICA method are found in functional networks. In reality, a few of ICs were background noise with few voxels that were triggered. Typically, functional networks have 1200 to 3000 active voxels. Fewer activated voxels were discovered to be just noise and not connection networks of relevance during the manual analysis by both professional reviewers. To decrease the number of ICs and enhance the effectiveness of functional network identification, an additional preliminary step was carried out to eliminate ICs from each patient with active voxels of fewer than 850. This threshold was set in order to define the bare minimum of active voxels above that may be regarded as a network. This threshold was designed to specify the bare minimum of active voxels over which a network can be considered. This approach proves critical in removing "noise" networks and enhance the sensitivity of our algorithm. Furthermore, before feeding the data into algorithms, we identified the coordinates of each cluster's greatest activation peak in order to reduce the number of variables evaluated for training.

Transfer learning strategies

We use SimCLR [19], a method based on contrastive learning, to efficiently learn visual representations from unlabeled images. Through a contrastive loss in a hidden representation of neural networks, SimCLR learns representations by maximizing agreement [20] between many augmented views of the same data sample. Given a mini-batch of images that were chosen at random, each image x_i is augmented twice using random rotation, Gaussian blur, and random crop resulting in two views of the same example both x_{2k1} and x_{2k}. To construct representations h_{2k1} and h_{2k}, the two images are encoded using an encoder network $f()$ (ResNet). After that, the representations are altered again using a non-linear transformation network $g()$, yielding z_{2k1} and z_{2k} that are used for the contrastive loss. The contrastive loss between two positive cases i, j (augmented from the same image) is presented using a mini-batch of encoded samples as follows

$$l_{i,j}^{NT-Xnet} = -log \frac{exp(sim(z_i, z_j)/\tau))}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} exp(sim(z_i, z_j)/\tau)} \quad (1)$$

where $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $[k \neq i]$, $sim(,)$ is cosine similarity between two vectors, and τ is a temperature scalar. We trained the model at a learning rate of $1e-5$ with $100,000$ epochs. To reduce model over-fitting, we adopt an early stopping method which uses the value of increase in validation error to make decision. Furthermore, we used grid-search algorithm to select optimal hyper-parameters for the SimCLR model related to increased precision of training data. The halting point of the training model was after $10,000$ validation failures and then a model checkpoint. For contrastive learning, we employed image augmentations including cropping, which pushes the model to encode various portions of the same image, as well as random translation, Gaussian blur, and random zoom layers. We concurrently loaded a large batch of unlabeled data from healthy subject images and a smaller batch of anotated samples from unhealthy subject images during training. We also used random horizontal flips as the second image augmentation method. To prevent overfitting on the few labeled samples, stronger augmentations, such cropping, are used for contrastive learning together with weaker ones, such horizontal flips, for supervised classification. The encoder model was pretrained on unanotated images with a defiend contrastive loss. The encoder's top is equipped with a nonlinear projection head, which enhances the quality of encoder representations. We employed the NT-Xent loss (Normalized Temperature-scaled Cross Entropy), which has the following meaning: Each image in the batch is treated as if it were its own class. Then, for each "class," we have two instances (a pair of augmented views). The representation of each perspective is compared to the representation of every possible pair (for both augmented versions). As logits, we employ the temperature-scaled cosine similarity of comparing representations. Finally, as the "classification" loss, we employ categorical cross-entropy. In order to monitor the pretraining performance, we used two metrics of contrastive accuracy [19] and linear probing accuracy. We fine-tuned the encoder on the annotated subjects, by adding

a single, fully connected classification layer with a random initialization on top.

III. Experimental Results and discussion

In this section, we provide details of outcomes from our experiments by using data collection procedure and training techniques explained in section II and recalled in 3. The values in table II display the experimental accuracy numbers that were recorded from different experiments organized from the adopted self-supervised model as well as a comparison with the proposed model in [4]. Data sizes that are utilized for testing and training were specified in each case. It is important to note that neither during training nor during hyper-parameter adjustment does the trained model ever view testing data.

TABLE II: Result of fMRI brain network classification with healthy and unhealthy data (7 fMRI network activation image corresponds to single patient in all cases).

	No. of training subjects	**No. of testing subjects**	**SimCLR**	**CNN** [4]
Healthy to Healthy	71 (labeled healthy)	10	81.74%	86%
Unhealthy to Unhealthy	45 (labeled unhealthy)	10	73.48%	75%
Healthy to Unhealthy	81 (labeled healthy)	55	69.21%	74%
Unhealthy to Unhealthy With unlabeled healthy	**81(unlabeled healthy) + 45 (labeled unhealthy)**	**10**	**76.39%**	—
Fine-tune on Unhealthy data from Healthy data	45 (labeled unhealthy)	10	—	78%

We performed data randomization at several points in the model training pipeline to provide a more consistent and reliable output, and we make sure the model has never seen test data before. Although the use of cross-validation techniques could be an alternative, we were unable to consider this option in order to keep our model simple and avoid further training complexity, which would have increased the computing resources needed for our contrastive learning model.

Initially, we trained and evaluated our model using data from healthy control subjects. This approach gave an absolute limit of performance with the highest accuracy of 81%, which is almost 5% less than the CNN model proposed in [4]. The accuracy evaluation in this case is very encouraging owing to the known spatial consistency in of healthy image data. In this experiment, the CNN model proposed in [4] reached the best performance on the CNN model compared with SimSLR. However, it has to be mentioned that a CNN method has a high cost of annotation of the data. In contrast, the self-supervised method can perform similarly with less annotated data.

In similar experiment, where training and testing of our model was organized with solely unhealthy patients, a reduction of about 8% compared to our previous result was recorded, which created a second baseline with fewer data. The same behavior was observed in [4] between the classification of unhealthy patients and healthy subjects, where a performance drop was around 11%. Although the performance on this baseline is less than the first experiment, the results are more important as this performance is achieved with fewer annotated unhealthy data, which is more interesting for clinical purposes.

On the brute-transfer strategy (learned from Healthy subjects without fine-tuning on unhealthy data), as shown in table II row 3, we trained both our self-supervised and supervised model with 81 annotated healthy control subjects and conducted testing on 55 unhealthy patients data. This time, we recorded an average accuracy of 69% for different ranges of test data sizes. It can be agreed that the brute-transfer learning does not introduce any accuracy enhancement in this case, similar to what was observed in [4]. Although, this observation highlights a significant difference between healthy and unhealthy patient data which quantifies its impact on transferability.

The fourth row of table II shows the performance of a new experiment where the SimCLR model is trained on a portion of unhealthy patients (45) and all unlabelled healthy data, which is fed to the model during the training among augmented images (pretext task). This experiment shows the most important result as its performance is more than the CNN model on unhealthy patients (2^{nd} row) and the brute-transfer learning (3^{rd} row) with about 3% and 7% respectively. The advantage of the SimCLR model in this experiment compared with other models in [4] is the use of non-labeled and few labeled data to train a model, while for CNN and transfer learning models, a large amount of label data is required.

The last row of the table II indicates the best performance of the transfer learning model in [4] while the CNN model has been trained once on all annotated healthy data. Then the model weights have been transferred and fined-tuned on unhealthy data. Although this model has the maximum accuracy among other experiments, the cost of the training model is too high as we need to use all 81 annotated healthy subjects and 45 annotated unhealthy subjects during the training. This cost can reduce the method's applicability for clinical purposes as we always lack annotated database in this domain, while the self-supervised method can gain similar performance with fewer data.

IV. Conclusion

An intriguing technique to accelerate learning in medical imaging, where healthy patients can be easily enrolled, is transfer learning with self-supervision. In contrast to traditional fine tuning following supervised learning, it does not require annotation of data from healthy subjects and data from unhealthy patients since there is no therapeutic interest. In this work, we provided an illustration on how to identify functional biological networks, and fortunately, the same method can be used for any non-invasive medical imaging task. Finally, these results further clarify the initial observation in the difference between healthy and unhealthy data.

Acknowledgment

LEI thanks Petroleum Technology Development Fund (PTDF)-NIGERIA.

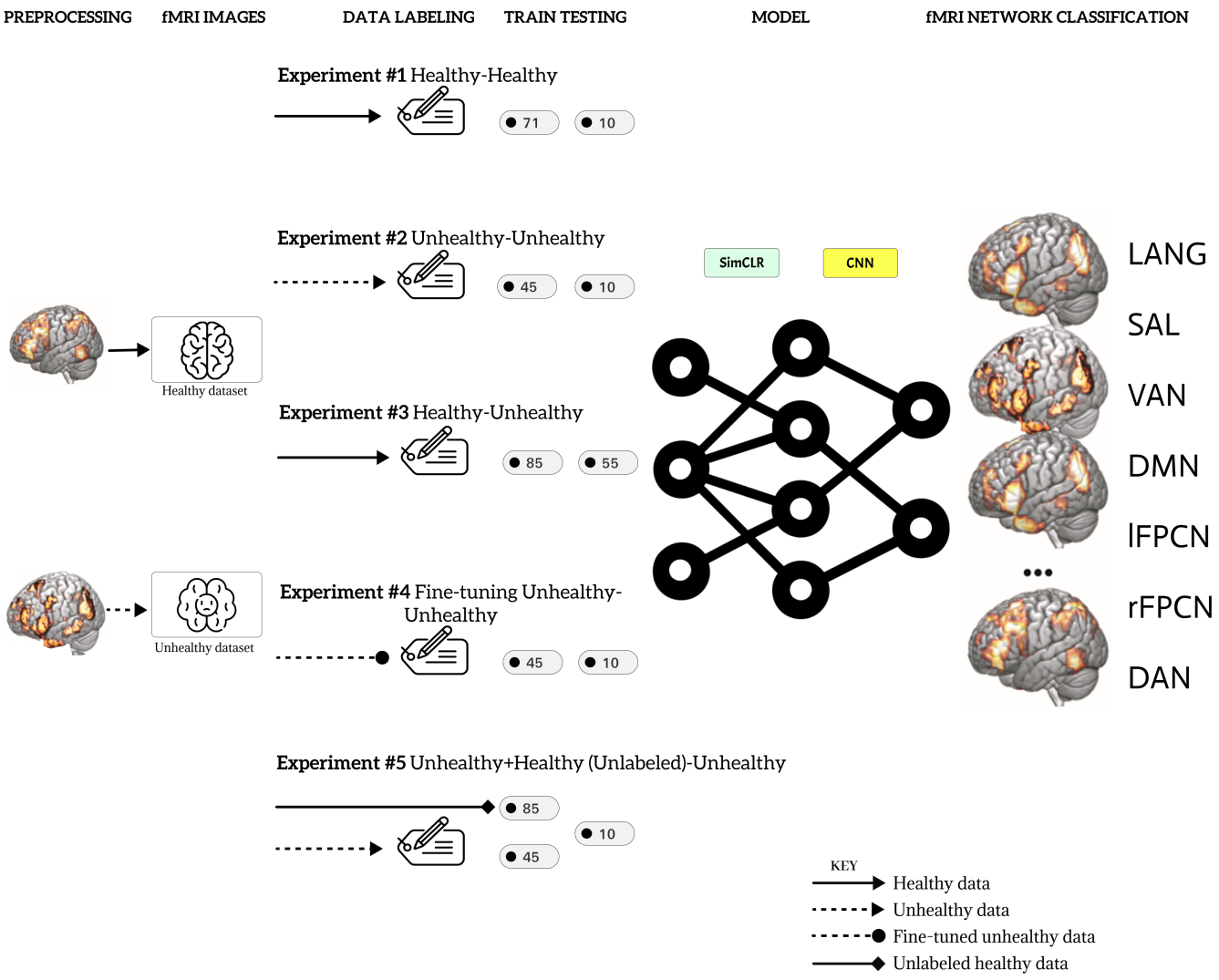


Fig. 3: Pipeline of Self-Supervised Learning with fMRI functional brain network classification of healthy and unhealthy image data.

REFERENCES

[1] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman *et al.*, "Artificial intelligence and machine learning for medical imaging: A technology review," *Physica Medica*, vol. 83, pp. 242–256, 2021.

[2] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," *Nature Biomedical Engineering*, pp. 1–16, 2022.

[3] Y. Liu, H. Zhang, W. Zhang, G. Lu, Q. Tian, and N. Ling, "Few-shot image classification: Current status and research trends," *Electronics*, vol. 11, no. 11, p. 1752, 2022.

[4] L. E. Ismaila, P. Rasti, F. Bernard, M. Labriffe, P. Menei, A. T. Minassian, D. Rousseau, and J.-M. Lemée, "Transfer learning from healthy to unhealthy patients for the automated classification of functional brain networks in fmri," *Applied Sciences*, vol. 12, no. 14, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/14/6925

[5] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, 2020.

[6] A. Jaiswal, A. R. Babu, M. Z. Zadeh, F. Makedon, and G. Wylie, "Understanding cognitive fatigue from fmri scans with self-supervised learning," *arXiv preprint arXiv:2106.15009*, 2021.

[7] I. Malkiel, G. Rosenman, L. Wolf, and T. Hendler, "Self-supervised transformers for fmri representation," in *Medical Imaging with Deep Learning*, 2021.

[8] J.-M. Lemée, D. H. Berro, F. Bernard, E. Chinier, L.-M. Leiber, P. Menei, and A. Ter Minassian, "Resting-state functional magnetic resonance imaging versus task-based activity for language mapping and correlation with perioperative cortical mapping," *Brain and behavior*, vol. 9, no. 10, p. e01362, 2019.

[9] Y.-O. Li, T. Adalı, and V. D. Calhoun, "Estimating the number of independent components for functional magnetic resonance imaging data," *Human brain mapping*, vol. 28, no. 11, pp. 1251–1266, 2007.

[10] H. I. Sair, N. Yahyavi-Firouz-Abadi, V. D. Calhoun, R. D. Airan,

S. Agarwal, J. Intrapiromkul, A. S. Choe, S. K. Gujar, B. Caffo, M. A. Lindquist *et al.*, "Presurgical brain mapping of the language network in patients with brain tumors using resting-state f mri: Comparison with task f mri," *Human brain mapping*, vol. 37, no. 3, pp. 913–923, 2016.

[11] F. Geranmayeh, R. J. Wise, A. Mehta, and R. Leech, "Overlapping networks engaged during spoken language production and its cognitive control," *Journal of Neuroscience*, vol. 34, no. 26, pp. 8728–8740, 2014.

[12] A. Ter Minassian, E. Ricalens, S. Nguyen The Tich, M. Dinomais, C. Aubé, and L. Beydon, "The presupplementary area within the language network: a resting state functional magnetic resonance imaging functional connectivity analysis," *Brain connectivity*, vol. 4, no. 6, pp. 440–453, 2014.

[13] C. Rosazza and L. Minati, "Resting-state brain networks: literature review and clinical applications," *Neurological sciences*, vol. 32, no. 5, pp. 773–785, 2011.

[14] M. H. Lee, C. D. Hacker, A. Z. Snyder, M. Corbetta, D. Zhang, E. C. Leuthardt, and J. S. Shimony, "Clustering of resting state networks," *PloS one*, vol. 7, no. 7, p. e40370, 2012.

[15] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster, "A probabilistic atlas of the human brain: Theory and rationale for its development: The international consortium for brain mapping (icbm)," *Neuroimage*, vol. 2, no. 2, pp. 89–101, 1995.

[16] D. Zhang, J. M. Johnston, M. D. Fox, E. C. Leuthardt, R. L. Grubb, M. R. Chicoine, M. D. Smyth, A. Z. Snyder, M. E. Raichle, and J. S. Shimony, "Preoperative sensorimotor mapping in brain tumor patients using spontaneous fluctuations in neuronal activity imaged with functional magnetic resonance imaging: initial experience," *Operative Neurosurgery*, vol. 65, no. suppl_6, pp. 226–236, 2009.

[17] C. F. Beckmann, M. DeLuca, J. T. Devlin, and S. M. Smith, "Investigations into resting-state connectivity using independent component analysis," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 1001–1013, 2005.

[18] B. R. Logan, M. P. Geliazkova, and D. B. Rowe, "An evaluation of spatial thresholding techniques in fmri analysis," *Human brain mapping*, vol. 29, no. 12, pp. 1379–1389, 2008.

[19] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[20] S. Becker and G. E. Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature*, vol. 355, no. 6356, pp. 161–163, 1992.