Explicability in resting-state fMRI for gender classification

Adrien Raison, Pascal Bourdon, Christophe N Habas, David Helbert

▶ **To cite this version:**

Adrien Raison, Pascal Bourdon, Christophe N Habas, David Helbert. Explicability in resting-state fMRI for gender classification. International Conference on Advances in Biomedical Engineering, Oct 2021, Wardanyeh, Lebanon. pp.5-8, 10.1109/ICABME53305.2021.9604842 . hal-03335712

HAL Id: hal-03335712
https://hal.science/hal-03335712

Submitted on 17 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Explicability in resting-state fMRI for gender classification

Adrien Raison[1,2,3], Pascal Bourdon[1,2,3], Christophe Habas[1,2,3,4,5], David Helbert[1,2,3]
[1] XLIM-ASALI, CNRS U-7252, University of Poitiers, France
[2] I3M Common Laboratory CNRS-Siemens, University and Hospital of Poitiers, France
[3] Poitiers University Hospital, CHU; Poitiers, France
[4] Neuroimaging Departement, Quinze Vingts Hospital, Paris, France
[5] University of Versailles Saint-Quentin, Versailles, France

November 16, 2021

Abstract

Artificial Intelligence, especially deep neural networks, have shown impressive performances for classification tasks since the last decade. In the medical field, trustworthy deep models exist but they do not provide any insights on how and why they classify data due to their complex structure. In this study we propose to leverage the power of deep neural network for classifying resting state brain activities by gender, then we use explainable Artificial Intelligence models to determine which functional networks are salient with respect to the gender. Firstly, we trained an accurate convolutional neural network to determine gender based on resting-state brain spatial maps corresponding to intrinsically connected networks and computed by independent component analysis. Then, we compare, through mask-based assessment, state of the art explainable Artificial Intelligence models to extract the most meaningful components involved in gender determination. Based on a powerful deep classifier, and with an appropriate explainable artificial intelligence method, we supply meaningful results in accordance with neurology literature results for gender classification. Throughout this study, we show that powerful deep models can be used in medical diagnostics since they recover, thank to reliable explainable artificial intelligence models, already established literature results related to gender determination with respect to brain network activities.

resting-state, functional magnetic resonance imaging, artificial intelligence, explainable artificial intelligence, gender classification

1 Introduction

Artificial Intelligence (AI) performs in humans feasible tasks where classical algorithms often fail. AI is based on a canonical functional representation:

Deep Neural Networks (DNN). Its complex structure allows it to permeate strong non-linearities involved in data describing daily life tasks. However, there is an inner back and forth communicating process related to performances and model opacities between regular and deep models. For classification tasks, regular algorithms are simpler but highly human-interpretable in terms of internal decision making processes, on the other side, DNN are highly performant but have very cloudy internals, often qualified as "black-box" models. A taxonomy of methods, labelled as eXplainable AI (XAI) methods, has been proposed by the machine learning community to explain DNN decision making processes. Functional neuroimaging, especially functional MRI (fMRI), permits to record in vivo spontaneous task-free or task-based brain activities. Deep models prove to be numerically accurate when applied to medical field, including the fMRI domain [13], [11], [17]. Deep learning has been applied to static and dynamic resting-state Functional Connectivity (FC) data for gender classification. But the lack of interpretability of these deep models constitutes a major flaw when they underlie medical diagnoses. FC refers to intrinsically connected networks, such as sensorimotor, limbic salience, central executive, default-mode, ventral and dorsal attentional and language-dedicated networks, which dynamically interact during task-free or task-positive brain activities. These networks are usually extracted from the global brain signal recorded with fMRI or electroencephalograms (EEG) using seed-based correlational or independent component analyzes (ICA). Several studies applying machine learning algorithms (SVM, GCN, LSTM-based...) to FC demonstrated that some networks, such as DMN, as well as their dynamic interaction allowed for gender identification. In this study, we aimed at: 1 leveraging high performance of DNN to achieve gender classification and 2. determining meaningful statistical dynamics embedding resting-state brain activity, using our masked-based XAI pipeline. For reproductibility purposes, we choose to use the publicly-available S1200 Wu-Minn Human Connectome Project to lead our study.

2 Problem Formulation

2.1 Data set

The data we use in this study is $N = 812$ subject resting state fMRI connectome data from the S1200 WU-Minn Human Connectome Project (released in June 2017 named as HCP1200 Parcellation+Timeseries+Netmats). The data consist of $T = 1,200$ volumes for each run for a total of 4,800 volumes for each subject over the four 15 minutes long runs. Each run of each subject's rs-fMRI was preprocessed by the HCP consortium [16]. The data were minimally preprocessed [6] and artefacts were removed using ICA+FIX [9] and [14]. We choose relatively low model order ICA (number of compo-

	22-25 y.	26-30 y.	31-35 y.	36+ y.	Total
Women	70	181	153	6	410
Men	125	176	99	2	402
Total	195	357	252	8	812

Table 1: Dataset subject distribution

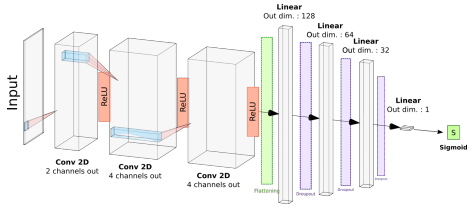

Figure 1: Model architecture

nents, $C = 25$) as previous studies have demonstrated that such models yield refined components that correspond to known anatomical and functional segmentations according to [2], [10], [16], [21]. The study subject demographic are shown in Table 1.

2.2 Procedure

Classification task Independent component analysis (ICA) generates statistically mutually independent spatial maps whose (neural) nodes display synchronized spontaneous BOLD fluctuations during the brain resting-state. These maps represent not only intrinsically connected networks but also noise due due to breathing, eye or head movements and spinal fluid pulsations easily identifiable by visual inspection or trained algorithms. These artifacts are indentified and discarded by [9]. To accurately classify subjects according to their gender and to account for sex differences in functional brain organization in a spatio-temporal manner, we train a convolutional neural network over 4 runs-averaged Independant Component (IC) time series. The network is composed of 3 convolutional blocks followed by 3 linear blocks. Each convolutional block is followed by a ReLU activation map, and each linear block has a dropout process enabled. Then, we apply the sigmoid mapping to obtain a value that leans in the [0,1] interval (see Figure 1). We use the ADAM version of stochastic gradient descent algorithm as optimization algorithm and we use the Binary Cross Entropy (BCE) as loss function.

2.2.1 XAI methods assessment

To compare the consistency of each XAI method to extract gender-oriented saliency maps (SM), we propose to assess each method throughout the whole

dataset by our normalized-mask process. We recall that $T = 1200$ is the number of brain snapshots for each IC and $C = 25$ is the number of ICs. We denote by $f : \mathbb{R}^{T\times C} \to [0, 1]$ our deep classifier and $g_f^m : \mathbb{R}^{T\times C} \to \mathbb{R}^{T\times C}$ the saliency mapping derived from f thanks to the explainer m. For a given explainer m, a classifier f and for each subject data $\mathbf{X}$, we obtained an explanation map $\widetilde{g_f^m}$ of the $\mathbf{X}$ classification by :

$$\widetilde{g_f^m}(\mathbf{X}) = s(g_f^m(\mathbf{X})) \odot \mathbf{X}$$

where s is a [0,1]-standardization mapping, such that for all well defined $\mathbf{A} \in \mathbb{R}^{T\times C}$:

$$s(\mathbf{A}) = \frac{\mathbf{A} - \min(\mathbf{A})}{\max(\mathbf{A}) - \min(\mathbf{A})} \in [0, 1]^{T\times C}$$

and where $\odot$ denotes the Hadamard product. The insights behind $\widetilde{g_f^m}$ transformation is to allow methods to be comparable since they don't produce the same output range thanks to standardization process as well as let them throw their own expressiveness capacities throught SMs. A XAI method is a method that redistribute as relevant as possible the information flow conditionally as the classification output. The literature already provides gender discriminating brain networks such as cingulate cortex, medial and lateral frontal cortex, temporoparietal regions, insula, and precuneus. For gender classification, we expect that XAI method provide larger importance weights to the aforementioned brain networks. The Pearson correlation study is a well-kwown statistical pipeline that highlights such networks interaction according gender and ages [4], [19], [7]. We denote the Pearson correlation matrix of an input $\mathbf{X}$ over the temporal axis as $PC(\mathbf{X}) \in \mathcal{M}_C([-1, 1])$. For ordering methods, we compare the Frobenius norm distribution of the matrix $PC(\mathbf{X}) - PC(\widetilde{g_f^m}(\mathbf{X})))$ throughout the whole dataset.

2.3 Interpretating methods

In this study, we compare the relevance of several state of the art XAI methods. These methods are qualified as local since they supply one explanation map by input and as intrinsic because the built explanation is network-dependant (architecture and weights).

Saliency (SA) [15] is a simple approach for computing input attribution, returning the gradient of the output with respect to the input. This approach can be understood as taking a first-order Taylor expansion of the network at the input, and the gradients are simply the coefficients of each feature in the linear representation of the model. The absolute value of these coefficients can be taken to represent feature importance.

Input $\times$ Gradient (IXG) is an extension of the Saliency approach, taking the gradients of the output with respect to the input and multiplying by the input feature values. One intuition for this approach considers a linear model; the gradients are simply the coefficients of each input, and the product of the input with a coefficient corresponds to the total contribution of the feature to the linear model's output.

Feature Ablation (FA) [12] is a perturbation based approach to compute attribution, involving replacing each input feature with a given baseline / reference value (e.g. 0), and computing the difference in output. Input features can also be grouped and ablated together rather than individually.

Feature Permutation (FP) [12] is a perturbation based approach which takes each feature individually, randomly permutes the feature values within a batch and computes the change in the loss as a result of this modification.

Layer Wise Relevant Propagation (LRP) [1] is an equally distributed feature relevance backpropagation system. The LRP method has been extensively used by the deep learning community especially in computer vision.

Deconvolution (DEC) [22] computes the gradient of the target output with respect to the input, but backpropagation of ReLU functions is overridden so that only non-negative gradients are backpropagated. In Deconvolution, the ReLU function is applied to the output gradients and directly backpropagated.

Score-CAM (SCAM) [18] is a post-hoc visual explanation method based on class activation mapping (CAM) [23], Score-CAM gets rid of the dependence on gradients by obtaining the weight of each activation map through its forward passing score on target class, the final result is obtained by a linear combination of these weights and upsampled activation maps of the lastest convolutional layer.

2.4 Network-IC correspondance mapping

For the given resolution C, we have constructed the correspondance mapping between known anatomo-functionnal networks and our C ICs. This mapping baseline has been handcrafted by an expert. The baseline mapping is summarized in Table 2.4.

IC	Network Name
1	Bilateral Visual Cortex (BVC)
2	Internal Visual Cortex (IVC)
3	Dorsal Attention Network (DAN)
4	Default Mode Network (DMN)
5	Right Central Executive Network (RCEN)
6	Salience Network (SN)
7	Left Central Executive Network (LCEN)
8	Precuneus / Default Mode Network (P/DMN)
9	Posterior Visual Network (PVN)
10	Bilateral Central Executive Network (BCEN)
11	Temporal Posterior Network (TPN)
12	Neocerebellum (N)
13	Motor Cortex (MC)
17	Frontal Singular Network (FSN)
18	Left Cerebellum (LC)
19	Precuneus (P)
20	Ventral Attention Network (VAN)
23	Fronto-percular Network (FPN)
25	Ventral Striatum (VS)

Table 2: Correspondance table between ICs and anatomo-functionnal networks. The components n°14,15,16,21,22,24 are assimilated as noisy components

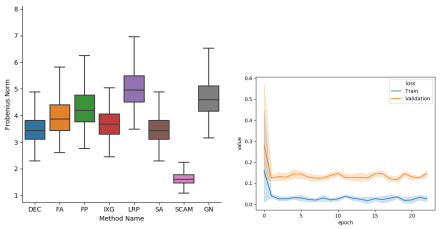

Figure 2: *Left* - Methods dispersion according to our pipeline. The box shows the quartiles of the Frobenius norm distributions derived from our evaluation. *Right* - Model classification performances.The line shows the mean loss for each epoch and the colored area show the 95% confidence interval of the estimated mean over 5 independents trainings.

3 Experimental Results

3.1 Classification task

Our trained model reaches a negligeable test error surrounding the 0.1 value for the binary cross entropy (BCE) after 22 epochs. We note that the learning pass is quiet stable along epochs.

3.2 Methods classification

Applied to fMRI data, XAI methods must at least recover the intrinsic functionnal activity present in the data. According to our pipeline, we observe in Figure 2 that Deconvolution, Feature Ablation, Feature Permutation, Input $\times$ Gradient and Saliency methods are clustered in neighborhood with respect to the Frobenius norm which is far away from the ICs temporal activity ground truth. That means that these methods does not recover the temporal linear interaction underlied in raw IC components. The Layer Wise Relevant Propagation is the farthest method and it is comparable to masked raw ICs time series with standard gaussian noise (GN) with respect to our process. This trend of providing poor SMs has been deeply investigated by [3]. Perturbation based methods do not provide any forward step towards better results for recovering ICs temporal activities. The highest ICs temporal activity fidelity is reach by the SCAM method.

3.3 Gender discrimination throught XAI methods

Literature always requires expert annotations to identify anatomo-functional networks across brains. Thanks to our DNN-based high classification score and the use of an adapted XAI method, we emulate the expert requirement to extract and recover gender discriminating networks. Studies [8], [20], [7], [19], [5] indicate that the strongest gender discriminating networks are the

Default Mode Network (IC n°4), Salience Network (IC n°6), the Left Central Executive Network (IC n°7) as well as the Right Central Executive Network (IC n°5). We address the gender comparing studies by gathering and temporally concatenating ICs activities respectively with respect to women and men. Then we compute, for each gender, the batch Pearson correlation matrices according to the SCAM method or LRP method.

For highlighting gender differences intrinsically present between women and men, we evaluate correlation differences amplitude between these two genders (see Figure 3). Not surprisingly we notice that there is difference between women and men mainly for the Default Mode Network, Salience Network, the Left/Right Central Executive Networks. Since the LRP method is irrelevant for our task, it does not rise up any meaningful differences between the two genders.

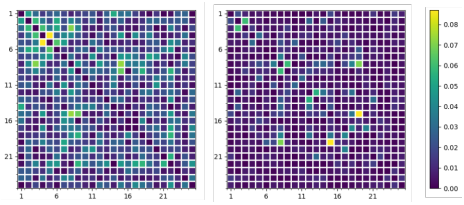

Figure 3: Correlation differences amplitude between Men-Women Pearson correlations matrix - Left : throught SCAM, Right : throught LRP

4 Discussion and Conclusion

Many regular algorithms for gender classification based on resting state brain networks activities have been proposed by the state of the art. Their high interpretability aspect is often in contrast with their performances. For diagnostics purposes, an interpretable model is more suitable than an opaque one. To improve medical diagnostics, we use the power of AI and apply SCAM model to grab classification decisions rules of our AI model. Our results show that such AI follow the same decision rules as regular model but with increased classification performance while recovering neurology literature results. This work is a preliminary work for studying newer XAI method over the same data or within an other paradigm such as task-related brain networks activities, or non healthy patient.

5 Acknowledgments

Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support

the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

References

[1] On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation.

[2] A. Abou-Elseoud, T. Starck, J. Remes, J. Nikkinen, O. Tervonen, and V. Kiviniemi. The effect of model order selection in group PICA. *Human Brain Mapping*, 31(8):1207–1216, Aug. 2010.

[3] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity Checks for Saliency Maps. *arXiv:1810.03292 [cs, stat]*, Nov. 2020. arXiv: 1810.03292.

[4] E. A. Allen, E. B. Erhardt, E. Damaraju, W. Gruner, J. M. Segall, R. F. Silva, M. Havlicek, S. Rachakonda, J. Fries, R. Kalyanam, A. M. Michael, A. Caprihan, J. A. Turner, T. Eichele, S. Adelsheim, A. D. Bryan, J. Bustillo, V. P. Clark, S. W. Feldstein Ewing, F. Filbey, C. C. Ford, K. Hutchison, R. E. Jung, K. A. Kiehl, P. Kodituwakku, Y. M. Komesu, A. R. Mayer, G. D. Pearlson, J. P. Phillips, J. R. Sadek, M. Stevens, U. Teuscher, R. J. Thoma, and V. D. Calhoun. A Baseline for the Multivariate Comparison of Resting-State Networks. *Frontiers in Systems Neuroscience*, 5, 2011. Publisher: Frontiers.

[5] E. Dhamala, K. W. Jamison, M. R. Sabuncu, and A. Kuceyeski. Sex classification using long-range temporal dependence of resting-state functional MRI time series. *Human Brain Mapping*, 41(13):3567–3579, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.25030.

[6] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson. The Minimal Preprocessing Pipelines for the Human Connectome Project. *NeuroImage*, 80:105–124, Oct. 2013.

[7] A. Goldstone, S. D. Mayhew, I. Przezdzik, R. S. Wilson, J. R. Hale, and A. P. Bagshaw. Gender Specific Re-organization of Resting-State Networks in Older Age. *Frontiers in Aging Neuroscience*, 8, 2016. Publisher: Frontiers.

[8] G. Gong, P. Rosa-Neto, F. Carbonell, Z. J. Chen, Y. He, and A. C. Evans. Age- and Gender-Related Differences in the Cortical Anatomical Network. *Journal of Neuroscience*, 29(50):15684–15693, Dec. 2009. Publisher: Society for Neuroscience Section: Articles.

[9] L. Griffanti, G. Salimi-Khorshidi, C. F. Beckmann, E. J. Auerbach, G. Douaud, C. E. Sexton, E. Zsoldos, K. P. Ebmeier, N. Filippini, C. E. Mackay, S. Moeller, J. Xu, E. Yacoub, G. Baselli, K. Ugurbil, K. L. Miller, and S. M. Smith. ICA-based artefact and accelerated fMRI acquisition for improved Resting State Network imaging. *NeuroImage*, 95:232–247, July 2014.

[10] V. Kiviniemi, T. Starck, J. Remes, X. Long, J. Nikkinen, M. Haapea, J. Veijola, I. Moilanen, M. Isohanni, Y.-F. Zang, and O. Tervonen. Functional segmentation of the brain cortex using high model order group PICA. *Human Brain Mapping*, 30(12):3865–3886, Dec. 2009.

[11] R. J. Meszlényi, K. Buza, and Z. Vidnyánszky. Resting State fMRI Functional Connectivity-Based Classification Using a Convolutional Neural Network Architecture. *Frontiers in Neuroinformatics*, 11, 2017. Publisher: Frontiers.

[12] C. Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

[13] P. Patel, P. Aggarwal, and A. Gupta. Classification of Schizophrenia versus normal subjects using deep learning. In *Proceedings of the Tenth Indian Conference on Computer Vision, Graphics and Image Processing*, ICVGIP '16, pages 1–6, New York, NY, USA, Dec. 2016. Association for Computing Machinery.

[14] G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith. Automatic Denoising of Functional MRI Data: Combining Independent Component Analysis and Hierarchical Fusion of Classifiers. *NeuroImage*, 90:449–468, Apr. 2014.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]*, Apr. 2014. arXiv: 1312.6034.

[16] S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, M. Kelly, T. Laumann, K. L. Miller, S. Moeller, S. Petersen, J. Power, G. Salimi-Khorshidi, A. Z. Snyder, A. T. Vu, M. W. Woolrich, J. Xu, E. Yacoub, K. Uğurbil, D. C. Van Essen, and M. F. Glasser. Resting-state fMRI in the Human Connectome Project. *NeuroImage*, 80:144–168, Oct. 2013.

[17] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *NeuroImage*, 129:292–307, Apr. 2016.

[18] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. *arXiv:1910.01279 [cs]*, Apr. 2020. arXiv: 1910.01279.

[19] S. Weis, K. R. Patil, F. Hoffstaedter, A. Nostro, B. T. T. Yeo, and S. B. Eickhoff. Sex Classification by Resting State Brain Connectivity. *Cerebral Cortex*, 30(2):824–835, Mar. 2020.

[20] C. Xu, C. Li, H. Wu, Y. Wu, S. Hu, Y. Zhu, W. Zhang, L. Wang, S. Zhu, J. Liu, Q. Zhang, J. Yang, and X. Zhang. Gender Differences in Cerebral Regional Homogeneity of Adult Healthy Volunteers: A Resting-State fMRI Study. *BioMed Research International*, 2015:1–8, 2015.

[21] M. Ystad, T. Eichele, A. J. Lundervold, and A. Lundervold. Subcortical functional connectivity and verbal episodic memory in healthy elderly–a resting state fMRI study. *NeuroImage*, 52(1):379–388, Aug. 2010.

[22] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]*, Nov. 2013. arXiv: 1311.2901.

[23] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *arXiv:1512.04150 [cs]*, Dec. 2015. arXiv: 1512.04150.