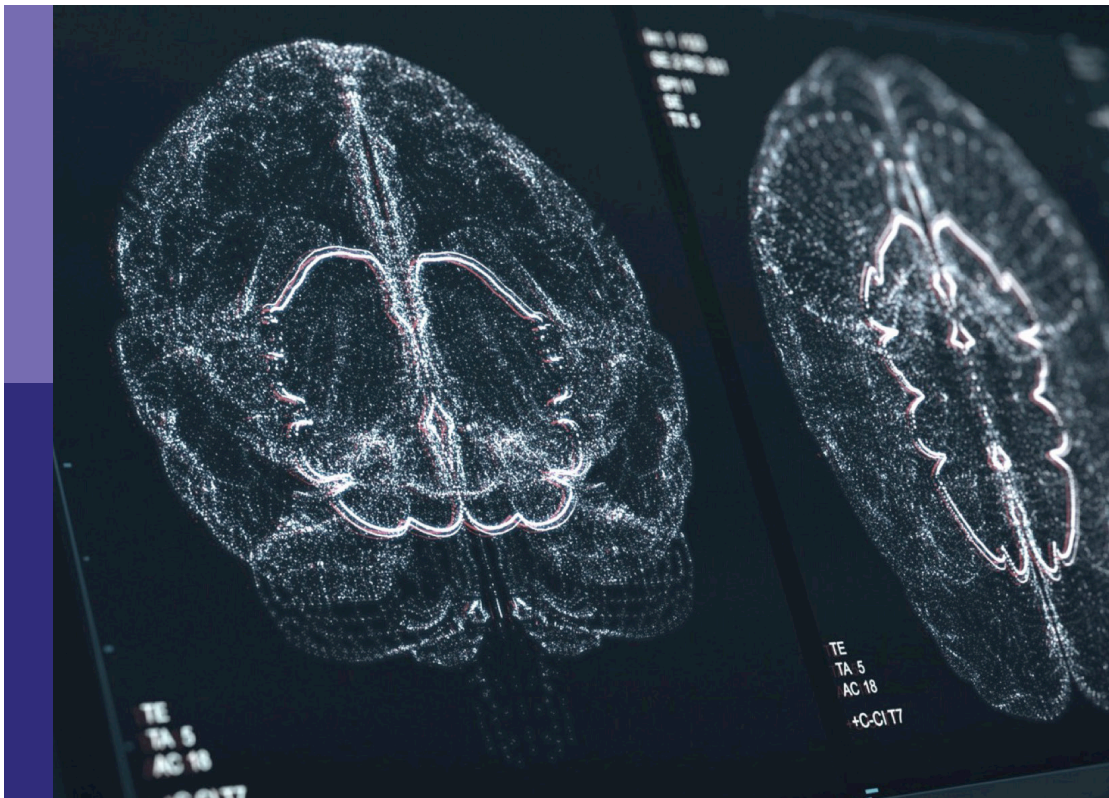
frontiers | Research Topics

Machine learning methods for human brain imaging

Edited by
Fatos Tunay Yarman Vural, Sharlene D. Newman, Tolga Cukur and Itir Onal Ertugrul

Published in
Frontiers in Neuroinformatics



ISSN 1664-8714
ISBN 978-2-83251-910-3
DOI 10.3389/978-2-83251-910-3

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Machine learning methods for human brain imaging

Topic editors

Fatos Tunay Yarman Vural — Middle East Technical University, Türkiye
Sharlene D. Newman — University of Alabama, United States
Tolga Cukur — Bilkent University, Türkiye
Itir Onal Ertugrul — Utrecht University, Netherlands

Citation

Vural, F. T. Y., Newman, S. D., Cukur, T., Ertugrul, I. O., eds. (2023). *Machine learning methods for human brain imaging*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83251-910-3

Table of contents

frontiers | Frontiers in Neuroinformatics

TYPE Editorial
PUBLISHED 28 February 2023
DOI 10.3389/fninf.2023.1154835

OPEN ACCESS

EDITED AND REVIEWED BY
Sean L. Hill,
Krembil Centre for Neuroinformatics, CAMH, Canada

*CORRESPONDENCE
Sharlene D. Newman
✉ sdnewman@ua.edu

RECEIVED 31 January 2023
ACCEPTED 10 February 2023
PUBLISHED 28 February 2023

CITATION
Yarman Vural FT, Newman SD, Çukur T and Önal Ertugrul I (2023) Editorial: Machine learning methods for human brain imaging. *Front. Neuroinform.* 17:1154835. doi: 10.3389/fninf.2023.1154835

Editorial: Machine learning methods for human brain imaging

Fatos Tunay Yarman Vural[1], Sharlene D. Newman[2]*, Tolga Çukur[3] and Itır Önal Ertugrul[4]

[1]Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, [2]Alabama Life Research Institute, The University of Alabama, Tuscaloosa, IN, United States, [3]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey, [4]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands

KEYWORDS

machine learning, deep learning, artificial intelligence, imaging, MRI

Editorial on the Research Topic
Machine learning methods for human brain imaging

The use of artificial intelligence (AI) methods like machine learning (ML), including deep learning, to make sense of brain imaging data has exploded over the past 10 years. Some of the early work focused on classifying brain states measured with functional magnetic resonance imaging (Mitchell et al., 2004). Those studies were exciting and demonstrated the potential power of ML to classify brain states in a way that reveals something about human cognition. ML is used in multiple aspects of brain imaging including image acquisition, reconstruction, analysis, and reporting (Aggarwal et al., 2023). For example, there are numerous studies using ML to classify groups of patients to improve diagnosis of neurodevelopmental disorders (e.g., autism, Parlett-Pelleriti et al., 2022), psychological disorders (e.g., schizophrenia, Chilla et al., 2022; and depression, Bhadra and Kumar, 2022), the progression of dementia (Mirzaei and Adeli, 2022) and tumors (Soomro et al., 2022), among others. On the image analysis side, ML applications are numerous and include it being used to improve denoising of image data (Gregory et al., 2021) and image segmentation (Wang et al., 2020).

The Research Topic, "*Machine learning methods for human brain imaging*," is a small sampling of 11 research articles that demonstrate the use of ML in multiple contexts and with multiple imaging modalities. The Research Topic includes two manuscripts (Alchihabi et al.; Fang et al.) that take different approaches to understanding cognitive networks—one using multi-variate pattern dependencies between brain regions and another examining network dynamics during the execution of a task. There are also three studies designed to use AI to diagnose psychological disorders—one using MRI to diagnosis defiant disorders in children (Menon and Krishnamurthy), one using EEG to classify brain states in schizophrenia patients and healthy controls (Plechawska-Wójcik et al.) and another classifying patients with obsessive-compulsive disorder and controls (Luo et al.). A third group of studies use ML to address analytic issues including one developing an open access tool for whole brain segmentation (Manjón et al.) and volumetric analysis of large datasets, one using fuzzy neural networks to improve 2D to 3D image transformations (Tavoosi et al.), and registration of multimodal 2D coronal section images of gene expressions in the mouse brain (Krepl et al.).

One goal of AI is to create systems that function like the human brain (Hopgood, 2005). Current systems fall short and two of the manuscripts in this Research Topic attempt to address this issue (Matsui et al.; Zhang et al.). Deep learning systems, for example do a great job of mimicking human vision, to a point; their mapping from stimulus input to perceptual output are different with respect to adversarial images. Zhang et al. attempts to characterize the differences in how AI systems and human brains process these adversarial images by comparing artificial neural networks and human brain activation, using what is learned to improve AI performance.

The use of ML in human brain imaging is only expected to increase. The power of deep learning methods makes them attractive for analyzing the growing number of large, publicly available datasets. However, it is important to slow down to evaluate their efficacy as well as to evaluate their weaknesses. One such weakness is addressed in the manuscript by Varotto et al.—how to handle imbalanced datasets. Most large datasets do not have even distributions of minority populations (e.g., racial, socioeconomic, patient, etc.). This is only one such shortcoming that demonstrates the need for careful evaluation.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aggarwal, K., Jimeno, M. M., Ravi, K. S., Gonzalez, G., and Geethanath, S. (2023). Developing and deploying deep learning models in brain MRI: a review. *arXiv preprint arXiv:2301.01241*. doi: 10.48550/arXiv.2301.01241

Bhadra, S., and Kumar, C. J. (2022). An insight into diagnosis of depression using machine learning techniques: a systematic review. *Curr. Med. Res. Opin.* 38, 749–771. doi: 10.1080/03007995.2022.2038487

Chilla, G. S., Yeow, L. Y., Chew, Q. H., Sim, K., and Prakash, K. B. (2022). Machine learning classification of schizophrenia patients and healthy controls using diverse neuroanatomical markers and Ensemble methods. *Sci. Rep.* 12, 2755. doi: 10.1038/s41598-022-06651-4

Gregory, S., Cheng, H., Newman, S., and Gan, Y. (2021). "HydraNet: a multi-branch convolutional neural network architecture for MRI denoising," in *Medical Imaging 2021: Image Processing, Vol. 11596* (Washington, DC: SPIE), 881–889. doi: 10.1117/12.2582286

Hopgood, A. A. (2005). The state of artificial intelligence. *Adv. Comput.* 65, 1–75. doi: 10.1016/S0065-2458(05)65001-2

Mirzaei, G., and Adeli, H. (2022). Machine learning techniques for diagnosis of Alzheimer disease, mild cognitive disorder, and other types of dementia. *Biomed. Signal Process. Control* 72, 103293. doi: 10.1016/j.bspc.2021.103293

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., et al. (2004). Learning to decode cognitive states from brain images. *Mach. Learn.* 57, 145–175. doi: 10.1023/B:MACH.0000035475.85309.1b

Parlett-Pelleriti, C. M., Stevens, E., Dixon, D., and Linstead, E. J. (2022). Applications of unsupervised machine learning in autism spectrum disorder research: a review. *Rev. J. Autism Dev. Disord.* 1–16. doi: 10.1007/s40489-021-00299-y

Soomro, T. A., Zheng, L., Afifi, A. J., Ali, A., Soomro, S., Yin, M., et al. (2022). Image segmentation for MR brain tumor detection using machine learning: a review. *IEEE Rev. Biomed. Eng.* 16, 70–90. doi: 10.1109/RBME.2022.3185292

Wang, J., Cheng, H., and Newman, S. D. (2020). Sparse representation of DWI images for fully automated brain tissue segmentation. *J. Neurosci. Methods* 343, 108828. doi: 10.1016/j.jneumeth.2020.108828

frontiers in Neuroinformatics

METHODS
published: 28 July 2021
doi: 10.3389/fninf.2021.691918

Supervised Learning With Perceptual Similarity for Multimodal Gene Expression Registration of a Mouse Brain Atlas

Jan Krepl†, Francesco Casalegno†, Emilie Delattre, Csaba Erö, Huanxiang Lu, Daniel Keller, Dimitri Rodarie, Henry Markram and Felix Schürmann*

Blue Brain Project, Ecole polytechnique fédérale de Lausanne, Genève, Switzerland

The acquisition of high quality maps of gene expression in the rodent brain is of fundamental importance to the neuroscience community. The generation of such datasets relies on registering individual gene expression images to a reference volume, a task encumbered by the diversity of staining techniques employed, and by deformations and artifacts in the soft tissue. Recently, deep learning models have garnered particular interest as a viable alternative to traditional intensity-based algorithms for image registration. In this work, we propose a supervised learning model for general multimodal 2D registration tasks, trained with a perceptual similarity loss on a dataset labeled by a human expert and augmented by synthetic local deformations. We demonstrate the results of our approach on the Allen Mouse Brain Atlas (AMBA), comprising whole brain Nissl and gene expression stains. We show that our framework and design of the loss function result in accurate and smooth predictions. Our model is able to generalize to unseen gene expressions and coronal sections, outperforming traditional intensity-based approaches in aligning complex brain structures.

Keywords: multimodal image registration, perceptual similarity, gene expression brain atlas, Allen mouse brain atlas, non-rigid, machine learning, deep learning

OPEN ACCESS

Edited by:
Itir Onal Ertugrul,
Tilburg University, Netherlands

Reviewed by:
Alexander K. Kozlov,
Royal Institute of Technology, Sweden
Anan Li,
Huazhong University of Science and Technology, China

****Correspondence:***
Jan Krepl
jan.krepl@epfl.ch

†These authors have contributed equally to this work and share first authorship

Received: *07 April 2021*
Accepted: *02 July 2021*
Published: *28 July 2021*

Citation:
Krepl J, Casalegno F, Delattre E, Erö C, Lu H, Keller D, Rodarie D, Markram H and Schürmann F (2021) Supervised Learning With Perceptual Similarity for Multimodal Gene Expression Registration of a Mouse Brain Atlas. Front. Neuroinform. 15:691918. doi: 10.3389/fninf.2021.691918

1. INTRODUCTION

Mouse brain atlases are an essential tool used by neuroscientists to investigate relationships between structural and functional properties of different brain regions. The Allen Institute for Brain Science has produced a reference whole brain atlas, associated Nissl stains, and about 20,000 different gene expression atlases obtained using high-throughput *in situ* hybridization (ISH) techniques (Lein et al., 2007; Dong, 2008).

In order to utilize the information provided by the different markers, gene expressions must be aligned to the reference Nissl atlas, so that all the data can be put into a common coordinate system. To this end, the Allen Mouse Brain Atlas (AMBA) includes an alignment module, but this module is limited to non-deformable transformations (Sunkin et al., 2012). For this reason, previous works (Erö et al., 2018) have had to resort to a manual landmark-based non-rigid approach to correct inaccuracies. However, this solution is not scalable to the whole genomic database.

We can describe our problem in terms of image registration, whereby the goal is to identify a transformation that maps a moving image to a target reference image. Our task is made particularly

challenging by the multimodality of gene expressions with respect to reference Nissl stains and by several artifacts like air bubbles and tears in the brain tissue samples.

In this work, we propose a supervised deep learning framework that efficiently leverages labels provided by a trained expert to accurately register multimodal 2D coronal section images showing gene expression stains. Our approach offers novel contributions in the following aspects:

1. Our model achieves high accuracy and generalizes to new gene expressions and coronal sections. It therefore constitutes a valuable tool for the integration of gene expression brain atlases.
2. By training with a perceptual similarity loss, our model learns to produce smooth deformations without the need any parametric constraint or post-processing stage.

1.1. Related Work

There has been some research on registration of Allen Brain datasets. Notably, Xiong et al. (2018) proposed a similarity metric addressing such artifacts and used it to register slices to the reference Nissl volume. Andonian et al. (2019) utilized groupwise registration to create multiple templates that are in turn used for pairwise registration of slices.

Among traditional image registration methods, intensity-based schemes (Klein et al., 2009) such as Symmetric Normalization (SyN) (Avants et al., 2008) represent the most popular approach. They do not require ground truth and rely on maximizing a similarity metric between the reference and registered moving image. These methods usually provide accurate and diffeomorphic predictions. However, they are limited by runtime overhead due to their intrinsically iterative nature, and also require a careful choice of hyperparameters. In particular, in the case of multimodal images like ours, tuning the pre-processing stages and the choice of the similarity metric required several time consuming trial-and-error iterations. In contrast, the model we propose can be easily deployed and used *as-is*, without the need for any tuning.

To address the limitations of traditional intensity-based approaches like SyN, several deep learning solutions have been proposed. Many approaches, such as VoxelMorph (Dalca et al., 2018; Balakrishnan et al., 2019), focused on unsupervised registration of magnetic resonance volumes following a similar approach to intensity-based models. Even though these methods reduced the runtime of the registration process, they cannot yield an improvement in accuracy over intensity-based methods, since they seek to optimize the same loss function (Lee et al., 2019). Furthermore, VoxelMorph maximizes cross-correlation, which is effective on unimodal data like magnetic resonance volumes, but fails on our multimodal images.

Among supervised approaches, RegNet (Sokooti et al., 2017) minimized mean absolute error with respect to a ground truth displacement field without adding any penalty guaranteeing smooth transformations. Moreover, this approach relied on synthetic training data and is therefore necessarily limited to unimodal problems and is therefore not applicable to our data.

Another popular supervised model is SVF-Net (Rohé et al., 2017), which has the advantage of training the model on ground truth transformations derived from region segmentation. The framework is based on training a network to align the boundaries of a pre-defined region of interest, which is not suitable for our use case since the visible brain regions vary across coronal sections.

Finally, while our proposed model learns to predict smooth deformations solely through the usage perceptual loss, previous methods relied either on: (i) parametric approaches like B-splines (de Vos et al., 2017), which restrict the space of possible deformations; (ii) introducing an explicit penalty term in the loss function (Balakrishnan et al., 2019), which further increases the number of hyperparameters; or (iii) integrating a predicted velocity field (Dalca et al., 2018), which requires post-processing steps.

The idea of training a model for image regression with a perceptual loss that uses the features extracted by a pre-trained network was first introduced in Johnson et al. (2016). In that work, the authors tested the approach on style transfer and super-resolution problems and showed that training with this loss produced models that better predict complex features such as texture and sharpness. The intuition behind this work was confirmed by Zhang et al. (2018), which proved that, on a variety of image datasets, the perceptual loss outperforms classical metrics in terms of correlation with human judgement. Perceptual loss has since then been successfully applied to various image generation tasks. To name a few, Huang et al. (2018) improved their results on higher resolutions when working on image-to-image translation, while Li et al. (2020) obtained artifact reduction and structure preservation on image denoising tasks.

Compared to these previous works using perceptual similarity, our approach also relies on the perceptual loss in order to learn to predict outputs that preserve complex visual features of the ground-truth, namely the smoothness of the displacements. However, our approach introduces elements of novelty in that we compute perceptual loss on the components of the displacement vector field rather than on images, and moreover we apply this approach to a new task such as multimodal image registration.

2. MATERIALS AND METHODS

Given a reference image I_{ref} and a moving image I_{mov}, image registration is defined as the problem of finding a transformation ϕ such that the registered image $I_{\text{reg}} = I_{\text{mov}} \circ \phi$ is as similar as possible to the reference I_{ref}. In the following, we assume that our input consists of a pair of multimodal images $I_{\text{ref}}, I_{\text{mov}} \in \mathbb{R}^{H\times W\times C}$ (H=height, W=width, C=number of channels), and that the output we want to predict is a transformation represented by an array $\phi \in \mathbb{R}^{H\times W\times 2}$ such that for every pixel (x, y) in I_{ref}, $\phi(x, y) \in \mathbb{R}^2$ defines the corresponding position of that pixel in I_{mov}. Equivalently, one can predict the per-pixel displacement $u \in \mathbb{R}^{H\times W\times 2}$ such that $u(x, y) = \phi(x, y) - (x, y)$.

The method we propose is based on supervised learning, so we assume that we have access to training samples $(I_{\text{ref}}, I_{\text{mov}}, \phi)$

where the ground truth label ϕ is provided by a human expert. These labeled samples are used to train a neural network model as described in the remainder of this section.

All the relevant code can be found at https://github.com/BlueBrain/atlas_alignment.

2.1. Network Architecture

Registration methods can be classified based on the family of transformations considered for the predicted deformation $\hat{\phi}$. Our model predicts pixel-wise displacements $\hat{u}$, so that it is non-parametric and allows for elastic transformations. This represents a considerable advantage in terms of expressive power in contrast to parametric models, such as affine or thin plate spline methods.

Specifically, the architecture of the neural network we propose is shown in **Figure 1**. Our model consists of two modules, predicting an affine (global) transformation $\hat{\phi}^{\text{global}}$ and an elastic (local) deformation $\hat{\phi}^{\text{local}}$, respectively. Our final prediction is the composition of the two transformations $\hat{\phi} = \hat{\phi}^{\text{global}} \circ \hat{\phi}^{\text{local}}$.

Unlike many related works on medical image registration (Sokooti et al., 2017; Yang et al., 2017; Balakrishnan et al., 2018; Dalca et al., 2018), we do not assume that our inputs are pre-centered and rescaled. Consequently, we employ a global alignment module to simplify the registration of the local one.

The architecture of the global and local modules are inspired by the Spatial Transformer Network (Jaderberg et al., 2015) and VoxelMorph (Balakrishnan et al., 2018), respectively.

2.2. Loss Function

In the case of multimodal registration, measuring image similarity between reference I_{ref} and predicted registration $\hat{I}_{\text{reg}} = I_{\text{mov}} \circ \hat{\phi}$ without pre-processing may provide misleading information due to the different appearance of these images. Thanks to our supervised learning framework, we can instead directly compare predictions $\hat{u}$ and $\hat{I}_{\text{reg}}$ with ground truths u and I_{reg}, respectively.

We train our model using a loss function composed of three terms

$$L_{\text{tot}} = L_{\text{IE}} + L_{\text{EPE}} + L_{\text{LPIPS}}. \tag{1}$$

The loss term L_{IE} is an *image error* between the predicted registered image $\hat{I}_{\text{reg}} = I_{\text{mov}} \circ \phi$ and the ground truth I_{reg}. As the two images have the same modality, pre-processing is unnecessary, and we can simply take

$$L_{\text{IE}} = \|I_{\text{reg}} - \hat{I}_{\text{reg}}\|_2^2. \tag{2}$$

The second term L_{EPE} is the squared average *endpoint error*, which is commonly used as a metric for optical flow estimation (Zhu et al., 2017). We define this loss as

$$L_{\text{EPE}} = \left(\sum_{x=1}^{H} \sum_{y=1}^{W} \frac{\|u(x,y) - \hat{u}(x,y)\|_2}{HWT} \right)^2, \tag{3}$$

where T is a normalizing constant representing the average displacement size computed from training data (in our case, $T \approx 20$).

Note that L_{EPE} is a pixel-wise loss which does not take into account information from neighboring pixels. As a consequence, our model often predicted non-smooth fields $\hat{\phi}$ with a significant number of corrupted pixels, i.e., (x, y) where the Jacobian $J_{\hat{\phi}}(x, y) \in \mathbb{R}^{2\times 2}$ has a non-positive determinant. In order to teach the model to predict transformations with smooth texture as the ground truth, we introduce in our total loss L_{tot} the loss term L_{LPIPS} defined as the *Learned Perceptual Image Patch Similarity Loss* version 0.1 with VGG-lin configuration (see Zhang et al., 2018 for details), which allows us to generate deformations that are perceptually similar to ground truth labels, including smoothness properties. To this end, we view the x and y components of u as two images.

Unlike other traditional metrics, L_{LPIPS} not only compares pixel-wise differences but also extracts and compares feature maps using a pre-trained VGG (Simonyan and Zisserman, 2014) network and then computes differences between these deep features. As explained in section 1.1, perceptual similarity is known to be effective in a variety of tasks where complex image features such as texture or sharpness have to be preserved in the predictions. In our case, the two images we compare are the ground truth and the predicted transformation, and the qualitative feature traits we try to preserve by relying on L_{LPIPS} consist in the smoothness of the ground truth displacements. Our results, presented and discussed in section 3.2, confirm the validity of this idea.

Finally, inspired by Zhao et al. (2019), we train our model to predict not only how to register I_{mov} to I_{ref}, but also I_{ref} to I_{mov}. We therefore effectively use $L'_{\text{EPE}} = L_{\text{EPE}}(u, \hat{u}) + L_{\text{EPE}}(u^{-1}, \widehat{u^{-1}})$ and $L'_{\text{LPIPS}} = L_{\text{LPIPS}}(u, \hat{u}) + L_{\text{LPIPS}}(u^{-1}, \widehat{u^{-1}})$. Given the ground truth u, we compute u^{-1} numerically using *scattered data interpolation* (SDI) (Crum et al., 2007).

To minimize the loss function L_{tot} we apply the RMSProp (Root Mean Square Propagation) optimizer (Tieleman and Hinton, 2012) which has a learning rate $\eta = 10^{-3}$ and a forgetting factor $\gamma = 0.9$. Additionally, the batch size is set to 4 due to GPU memory limitations.

2.3. Data Augmentation

To improve the generalization performance of our model, we generate synthetic samples of the form $(I_{\text{ref}}, I'_{\text{mov}}, \phi')$ from each training sample $(I_{\text{ref}}, I_{\text{mov}}, \phi)$.

A first class of augmentations generates I'_{mov} from I_{mov} by applying random blurring, brightness perturbation, and other image processing techniques. These augmentations help improve the accuracy of predictions on images having different perceptual appearance, and generalize to gene expressions not present in the training set. Note that for these augmentations $\phi' = \phi$.

Another class of augmentations consists of geometric transformations, affecting both I_{mov} and ϕ. These are particularly relevant for our application, since our focus is on predicting elastic deformations. First, control points are sampled on the edges of a brain section, and random displacements are generated for each of these points. Interpolating these displacement vectors with radial basis functions yields a smooth transformation ψ

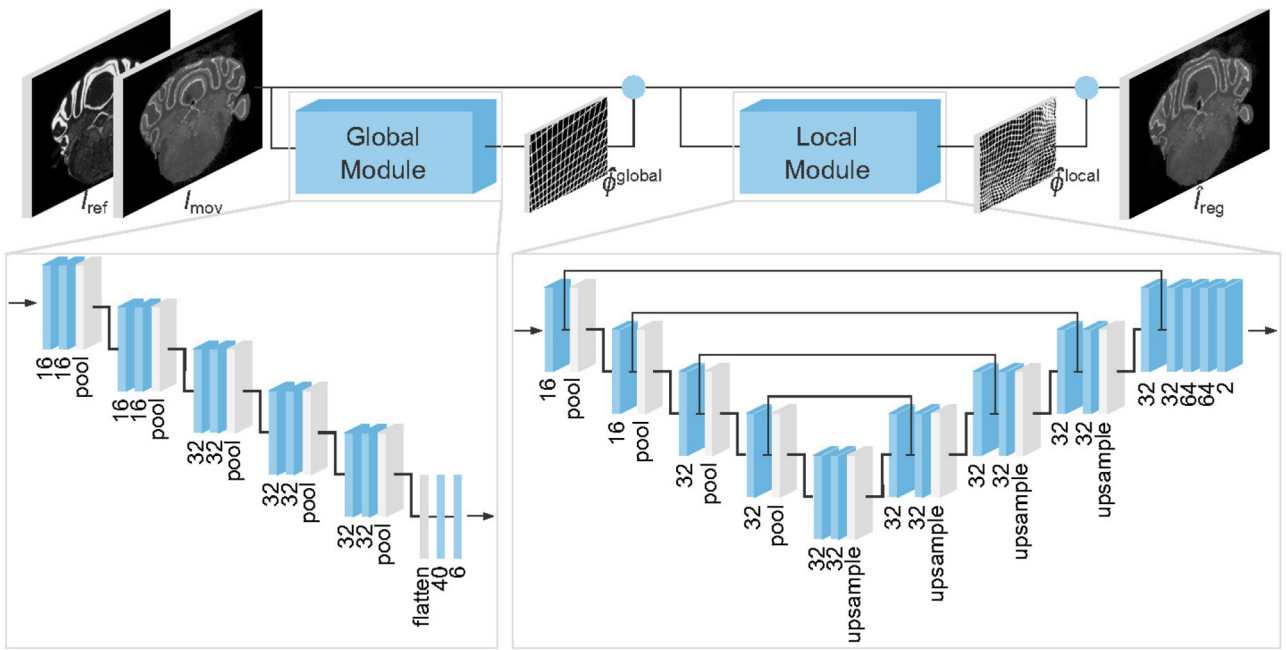


FIGURE 1 | Architecture of the proposed model to register gene expressions I_{mov} to Nissl stains I_{ref}. Blue boxes represent trainable layers. Under all 3×3 convolutional layers (3D boxes) and fully connected layers (2D boxes) we display the number of channels and dense units, respectively. Blue circles denote where predicted transformations $\hat{\phi}$ are applied.

defined over the whole I_{mov}. We obtain a synthetic sample by considering $I'_{mov} = I_{mov} \circ \psi$ and $\phi' = \phi \circ \psi^{-1}$.

2.4. Dataset and Evaluation Metrics

The reference Nissl stain volume of the AMBA comprises 528 coronal sections. Typically 8 markers per specimen were assayed, yielding approximately 60 coronal sections per gene expression. Our goal is to register the moving gene expression I_{mov} to the reference Nissl slice I_{ref}.

In order to train and evaluate our model, we selected 277 section pairs from the Nissl atlas and 7 different gene atlases for calbindin (CALB1), calretinin (CALB2), cholecystokinin (CCK), neuropeptide Y (NPY), parvalbumin (PVALB), somatostatin (SST), and vasointestinal peptide (VIP). Even though all the gene expressions were pre-aligned using the affine registration module provided by the Allen Brain API, significant misalignments were still present. The original sections have various resolutions, so we had to rescale the images in order to be able to run our model, which assumes all moving and reference inputs to have the same shape. We therefore downscaled all slices to a fixed 320×456 pixels resolution, which corresponds to a 25 µm sampling distance that is the same value of the slices thickness in the Nissl atlas, in order to have a uniform resolution across the three axes. Finally, all images were converted to grayscale.

We collected ground truth labels from a human expert provided with a manual landmark-based non-rigid registration tool that we designed to export the deformation field and registered image. This annotation tool is named `label-tool` and is part of our open-source Python package. On average the expert used 27.7 keypoint pairs (with a standard deviation of 10.7) to register a sample. However, the number required keypoints significantly depends on the gene expression and on the coronal section location, as shown in **Table 1**. This provides a further argument in favor of our supervised learning approach, which exports the whole deformation field provided by a human expert and does not constrain the annotator to a fixed number of control points, unlike the case of parameteric models such as de Vos et al. (2017).

TABLE 1 | Average number of keypoint pairs used by the annotation expert (per gene and coronal section group).

Gene	1–176	177–352	353–528
CALB1	19.1	36.3	40.8
CALB2	18.7	36.9	39.4
CCK	18.6	30.4	35.1
NPY	19.4	27.4	27.0
PVALB	17.9	24.4	37.3
SST	17.7	29.6	26.0
VIP	19.6	31.2	33.4

To measure the performance of our model, we considered the hierarchical segmentation maps provided by the AMBA to compute the average Dice score (Dice, 1945) (using weights proportional to the number of pixels of each segmentation class) at different levels, as shown in **Figure 2**. We performed this comparison in the moving space by warping the ground truth segmentation by ϕ^{-1} and $\hat{\phi}^{-1}$ (both computed numerically).

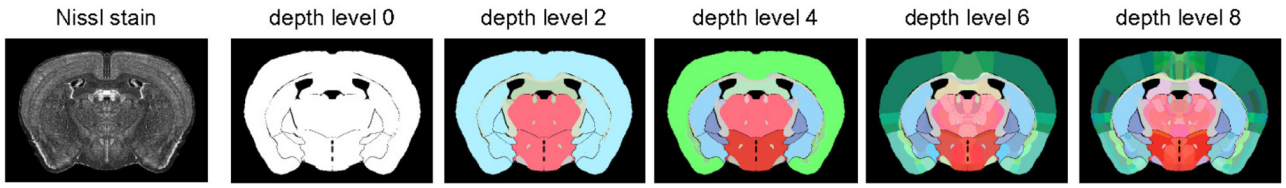


FIGURE 2 | Hierarchical segmentation of a Nissl stain (coronal section 250) used to compute Dice score to evaluate our model. Level 0 only distinguishes background from foreground, while deeper levels define an increasing number of brain subregions.

TABLE 2 | Summary of results on an 80:20 train-test stratified data split (mean and standard deviation in percentage).

Model	Dice-0	Dice-2	Dice-4	Dice-6	Dice-8	$\lvert J_{\hat{\phi}} \rvert \leq 0$
Ours	**94.2 ± 4.0**	**84.4 ± 6.9**	**80.6 ± 7.9**	**68.0 ± 13.3**	**55.2 ± 11.9**	0.11 ± 0.17
SyN	94.1 ± 4.2	83.9 ± 7.6	79.8 ± 8.8	66.1 ± 13.9	52.3 ± 12.5	0.01 ± 0.02
Affine	91.4 ± 5.9	79.9 ± 10.0	75.5 ± 11.0	61.2 ± 17.7	46.8 ± 15.8	0.00 ± 0.00

Bold values indicate the highest (= best) Dice score in the various experiments.

As a benchmark, we used an affine model and SyN as implemented in the Advanced Normalization Tools (ANTs) software package (Avants et al., 2011). We opted for mutual information as a similarity metric to handle multimodality.

3. RESULTS

3.1. Quantitative Analysis

We evaluated the performance of our model on two different experiments. In the first experiment, we applied an 80:20 train-test split using a stratified partitioning scheme based on the different genes and on the section locations on the anterior-posterior axis.

As indicated in **Table 2**, our model outperforms both the affine model and SyN with respect to Dice score. The improvement over SyN is marginal for level 0, which corresponds to a background-foreground segmentation as shown in **Figure 2**. However, our model's relative advantage increases as we consider more regions. Indeed, aligning complex brain structures in multimodal images is a harder task for intensity-based models. **Table 2** shows that our model tends to predict smooth transformations with only 0.11% of corrupted pixels, mostly occurring at image borders. This is particularly noteworthy since the smoothness emerges naturally from training with the loss function defined in section 2.2.

In the second experiment, we studied how our model generalizes to new genes by training on slices of 6 genes and evaluating performances on the remaining holdout gene. Results in terms of Dice-8 score, where difference between models is more visible, are reported in **Table 3**. Even in this more difficult scenario, where slices of the holdout gene are never shown to the model during the training phase, our network achieves higher scores than SyN on all but one gene. The overall results of this second experiment confirm that our model generalizes to new genes and is therefore suitable for registering and leveraging multimodal gene atlases.

Finally, running on an Intel Core i7-4770 CPU, registering a sample takes either $\sim$ 3 s or $\sim$ 0.2 s using SyN or our model, respectively. On an NVIDIA Tesla V100 GPU, the runtime of our model is further reduced to $\sim$ 0.009 s (the ANTs package does not provide GPU implementations of SyN). These results demonstrate that our approach is also competitive in terms of runtime.

3.2. Qualitative Analysis

A qualitative analysis of the predictions of our model is shown in **Figure 3**. Our global module provides a first affine transformation that rescales and centers the moving image. The need and the efficacy of this module are particularly visible in the case of samples (**Figures 3B,C**), where the global module significantly rescales and shifts the input gene expression. The local module then applies an elastic deformation that accurately aligns the gene expression to the reference Nissl stain.

We already mentioned in section 1 that our registration task is made particularly challenging by the presence of tears and air bubbles in the gene expression stains. In **Figure 4**, we demonstrate the stability of our approach by showing examples of gene expression slices including these kinds of artifacts together with the ground truth and predicted registrations.

As explained previously in section 2.2, the smoothness of the predicted deformation field $\hat{\phi}$ can be entirely ascribed to our choice of loss function. **Figure 5** illustrates how results vary depending on whether or not L_{tot} includes the perceptual similarity term L_{LPIPS}. Notice that, without this term, the model produces a significant number of corrupted pixels.

Further insight with respect to these results is provided in **Figure 6**, where we can observe some of the feature maps used to compute L_{LPIPS}. As previously described, these deep features are the internal activations of a pre-trained VGG network. The similar, smooth appearance of the ground truth u and predicted transformation $\hat{u}$ obtained by training with L_{LPIPS} is well-captured by these activations, which look significantly different

TABLE 3 | Summary of results on a gene-holdout split (Dice-8, mean, and standard deviation in percentage).

Model	CALB1	CALB2	CCK	NPY	PVALB	SST	VIP
Ours	**48.8 ± 9.5**	**55.3 ± 12.3**	**54.5 ± 13.0**	**44.4 ± 12.5**	56.0 ± 13.6	**60.3 ± 12.4**	**58.4 ± 13.3**
SyN	46.9 ± 10.9	54.6 ± 12.9	50.7 ± 14.3	41.7 ± 15.4	**58.5 ± 10.6**	57.0 ± 12.5	55.6 ± 11.5
Affine	40.5 ± 15.8	51.5 ± 14.3	46.7 ± 12.0	36.6 ± 15.5	53.0 ± 12.1	54.3 ± 13.4	52.5 ± 14.3

Bold values indicate the highest (= best) Dice score in the various experiments.

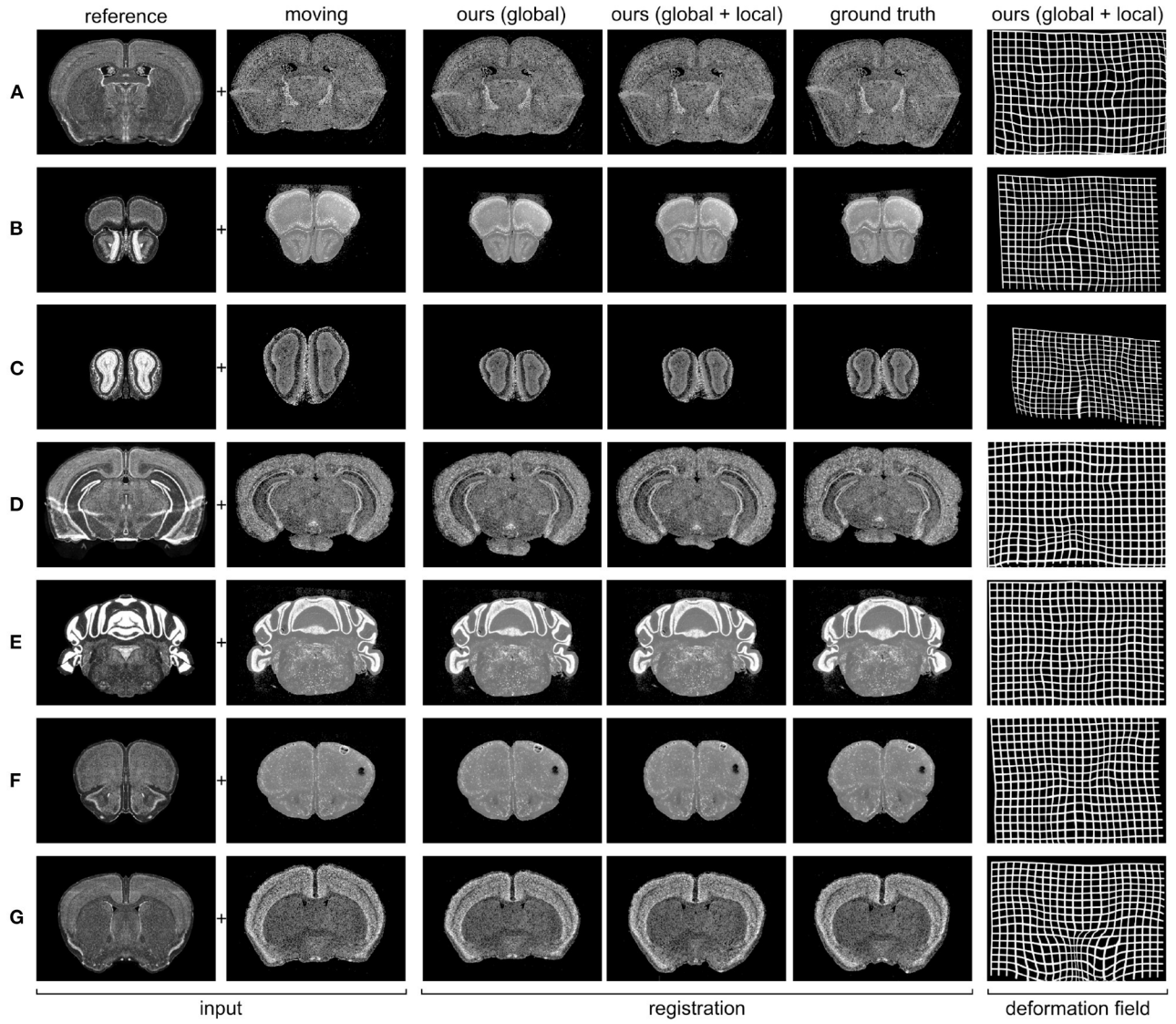


FIGURE 3 | Predicted registrations on slices from the 7 different gene expression atlases used in our experiments (see section 2.4 for details). **(A)** PVALB gene, section 236; **(B)** CALB1 gene, section 100; **(C)** NPY gene, section 52; **(D)** SST gene, section 328; **(E)** CALB2 gene, section 451; **(F)** VIP gene, section 129; **(G)** CCK gene, section 190.

for the non-smooth predicted transformation $\hat{u}$ we obtained when training without the L_{LPIPS} term. These observations help justify the importance of using the perceptual loss in our framework to produce smooth results.

Interestingly, if we evaluate the predicted transformations shown in **Figure 6** using L_{EPE}, the prediction obtained by training with the perceptual loss ($L_{\text{EPE}} = 6.83$) seems to be worse than the one obtained without it ($L_{\text{EPE}} = 6.03$). This strongly contrasts with the fact that this latter looks smooth and qualitatively similar to the ground truth, while the other prediction clearly includes a large number of artifacts. However, if we evaluate the same transformations using L_{LPIPS} we reach the opposite

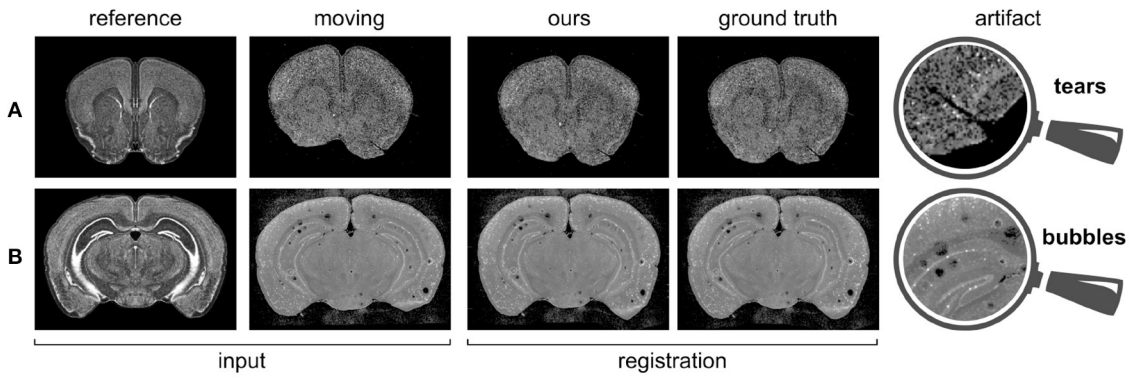


FIGURE 4 | Gene expressions containing artifacts, and corresponding predicted registrations. **(A)** PVALB gene, section 160; **(B)** VIP gene, section 316.

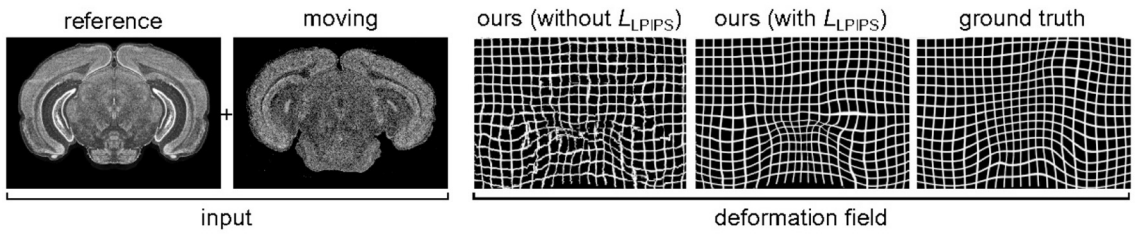


FIGURE 5 | Influence of the loss function on the smoothness of the predicted deformation, for SST gene, section 352. If we use a loss without perceptual similarity, $\sim 3\%$ of pixels are corrupted. By introducing the L_{LPIPS} term, this is reduced to $\sim 0.1\%$.

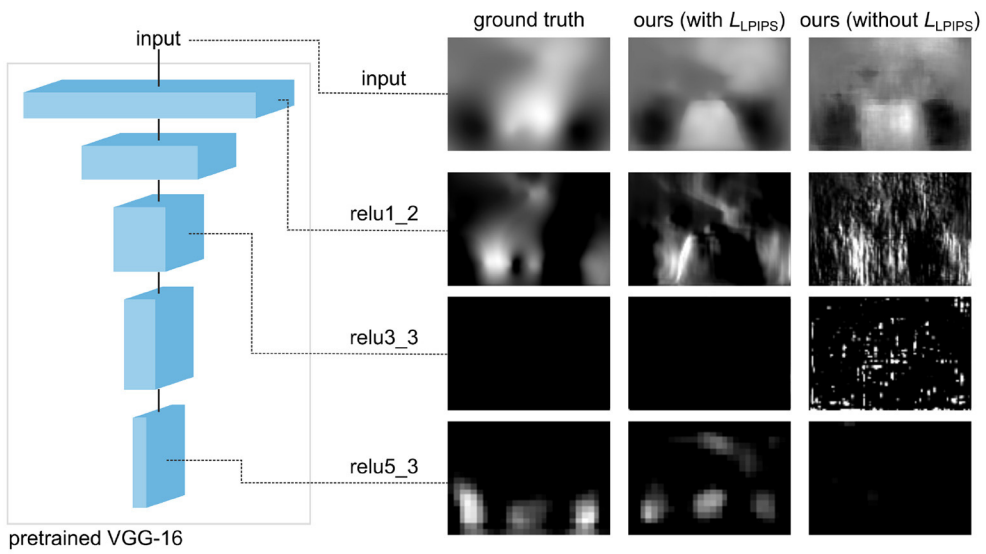


FIGURE 6 | Activations of the pre-trained network used to compute L_{LPIPS} on the y component of the transformation u for the coronal section 352 of SST gene (same as in **Figure 5**). The deep features of the non-smooth predicted $\hat{u}$ obtained by training without L_{LPIPS} significantly differ from those of smooth transformations corresponding to the ground truth u and predicted $\hat{u}$ obtained by training with L_{LPIPS}.

conclusions, as the prediction obtained by training with the perceptual loss ($L_{\text{LPIPS}} = 0.30$) appears to be better than the one obtained without it ($L_{\text{LPIPS}} = 0.53$). These results are consistent with Zhang et al. (2018), where perceptual similarity is shown to strongly correlate with human perception, unlike other traditional metrics.

4. DISCUSSION

In this paper, we presented a supervised deep learning model with perceptual similarity for the 2D registration of gene expressions to Nissl stains of the Allen Mouse Brain Atlas. The main novelty of our method lies in its unique non-parametric approach which allows the prediction of smooth deformations by exclusively relying on a perceptual loss function. In contrast to this, previous works had to resort to using parametric methods, extra penalty terms with hyperparameters requiring careful tuning, or post-processing steps.

By testing on two different experiments, we showed that the proposed approach produces accurate predictions that generalize well to unseen gene expressions and coronal sections. This is particularly significant given the high variability of shape and appearance across stains and sections, as shown in **Figure 3**. We benchmarked our results against the state-of-the-art method SyN, and our results showed that our model is significantly faster and it also achieves higher accuracy in almost all cases.

Our qualitative analysis shows that our model is able to predict deformation fields that are very close to the ground truth annotations provided by a human expert, even in case of slices affected by artifacts such as air bubbles and tears. Indeed, during the training phase, our model is presented with samples including various kinds of anomalies, and therefore learns how to predict a deformation field in a correct way, as opposed to intensity-based approaches.

Our framework has therefore proven capable of enabling the neuroscience community to leverage large-scale complex brain-derived datasets, with a significant scientific impact in terms of acceleration and accuracy improvement.

We identify three drawbacks of the presented approach. Firstly, it assumes that we have access to expert labels. Manual registration with any annotation tool is a difficult task and the resulting ground truth deformation might vary from one expert to another. The second shortcoming is that a generalization of our approach to 3D registration is not straightforward. This is due to the fact that perceptual loss is computed on images rather than volumes. Lastly, the training of our neural network represents the most time consuming stage of the pipeline. This is a common problem of many deep learning models and it should not be completely overshadowed by fast inference.

The future research direction is to apply our approach to new datasets. One specific example is to investigate sagittal sections. In general, the most promising applications are in the registration of multimodal datasets where using traditional approaches might lead to inaccurate results.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

JK and FC conceived and designed the method. JK, CE, HL, and DR collected the data and including the annotations. All authors contributed to interpreting the results and writing the paper, approved the final version, and agreed to be accountable for all aspects of the work.

FUNDING

This study was supported by funding to the Blue Brain Project, a research center of the Ecole polytechnique fédérale de Lausanne (EPFL), from the Swiss government's ETH Board of the Swiss Federal Institutes of Technology.

REFERENCES

Andonian, A., Paseltiner, D., Gould, T. J., and Castro, J. B. (2019). A deep learning based method for large-scale classification, registration, and clustering of *in-situ* hybridization experiments in the mouse olfactory bulb. *J. Neurosci. Methods* 312, 162–168. doi: 10.1016/j.jneumeth.2018.12.003

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., and Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* 54, 2033–2044. doi: 10.1016/j.neuroimage.2010.09.025

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2018). "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 9252–9260. doi: 10.1109/CVPR.2018.00964

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V. (2019). VoxelMorph: a learning framework for deformable medical image registration. *IEEE Trans. Med. Imaging* 38, 1788–1800. doi: 10.1109/TMI.2019.2897538

Crum, W. R., Camara, O., and Hawkes, D. J. (2007). "Methods for inverting dense displacement fields: evaluation in brain image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Brisbane, QLD: Springer), 900–907. doi: 10.1007/978-3-540-75757-3_109

Dalca, A. V., Balakrishnan, G., Guttag, J., and Sabuncu, M. R. (2018). "Unsupervised learning for fast probabilistic diffeomorphic registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Granada: Springer), 729–738. doi: 10.1007/978-3-030-00928-1_82

de Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M., and Išgum, I. (2017). "End-to-end unsupervised deformable image registration with a convolutional neural network," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Quebec City, QC: Springer), 204–212. doi: 10.1007/978-3-319-67558-9_24

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi: 10.2307/1932409

Dong, H. W. (2008). *The Allen Reference Atlas: A Digital Color Brain Atlas of the C57Bl/6J Male Mouse*. Hoboken, NJ: John Wiley & Sons Inc.

Erö, C., Gewaltig, M.-O., Keller, D., and Markram, H. (2018). A cell atlas for the mouse brain. *Front. Neuroinform.* 12:84. doi: 10.3389/fninf.2018.00084

Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. (2018). "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 172–189. doi: 10.1007/978-3-030-01219-9_11

Jaderberg, M., Simonyan, K., Zisserman, A., and Kavukcuoglu, K. (2015). Spatial transformer networks," in *Advances in Neural Information Processing Systems*, eds C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Montréal, QC: NIPS Proceedings) 2017–2025.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision* (Amsterdam: Springer), 694–711. doi: 10.1007/978-3-319-46475-6_43

Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. (2009). elastix: a toolbox for intensity-based medical image registration. *IEEE Trans. Med. Imaging* 29, 196–205. doi: 10.1109/TMI.2009.2035616

Lee, M. C., Oktay, O., Schuh, A., Schaap, M., and Glocker, B. (2019). "Image-and-spatial transformer networks for structure-guided image registration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen: Springer), 337–345. doi: 10.1007/978-3-030-32245-8_38

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168–176. doi: 10.1038/nature05453

Li, M., Hsu, W., Xie, X., Cong, J., and Gao, W. (2020). SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network. *IEEE Trans. Med. Imaging* 39, 2289–2301. doi: 10.1109/TMI.2020.2968472

Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M., and Pennec, X. (2017). "SVF-Net: learning deformable image registration using shape matching," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Quebec City, QC: Springer), 266–274. doi: 10.1007/978-3-319-66182-7_31

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. Available online at: https://arxiv.org/abs/1409.1556

Sokooti, H., De Vos, B., Berendsen, F., Lelieveldt, B. P., Išgum, I., and Staring, M. (2017). "Nonrigid image registration using multi-scale 3D convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Quebec City, QC: Springer), 232–239. doi: 10.1007/978-3-319-66182-7_27

Sunkin, S. M., Ng, L., Lau, C., Dolbeare, T., Gilbert, T. L., Thompson, C. L., et al. (2012). Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* 41, D996–D1008. doi: 10.1093/nar/gks1042

Tieleman, T., and Hinton, G. (2012). Lecture 6.5-RMSProp: divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw. Mach. Learn.* 4, 26–31. Available online at: https://www.reddit.com/r/MachineLearning/comments/2lmo0l/ama_geoffrey_hinton/clyjvky?utm_source=share&utm_medium=web2x&context=3

Xiong, J., Ren, J., Luo, L., and Horowitz, M. (2018). Mapping histological slice sequences to the Allen Mouse Brain Atlas without 3D reconstruction. *Front. Neuroinform.* 12:93. doi: 10.3389/fninf.2018.00093

Yang, X., Kwitt, R., Styner, M., and Niethammer, M. (2017). Quicksilver: fast predictive image registration-a deep learning approach. *Neuroimage* 158, 378–396. doi: 10.1016/j.neuroimage.2017.07.008

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 586–595. doi: 10.1109/CVPR.2018.00068

Zhao, S., Lau, T., Luo, J., Eric, I., Chang, C., and Xu, Y. (2019). Unsupervised 3D end-to-end medical image registration with volume tweening network. *IEEE J. Biomed. Health Inform* 24, 1394–1404. doi: 10.1109/JBHI.2019.2951024

Zhu, Y., Lan, Z., Newsam, S., and Hauptmann, A. G. (2017). Guided optical flow learning. *arXiv preprint arXiv:1702.02295*. Available online at: https://arxiv.org/abs/1702.02295

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fninf.2021.677925

Dissociable Neural Representations of Adversarially Perturbed Images in Convolutional Neural Networks and the Human Brain

Chi Zhang[1], Xiao-Han Duan[1], Lin-Yuan Wang[1], Yong-Li Li[2], Bin Yan[1], Guo-En Hu[1], Ru-Yuan Zhang[3,4]† and Li Tong[1]*†*

[1] Henan Key Laboratory of Imaging and Intelligent Processing, PLA Strategic Support Force Information Engineering University, Zhengzhou, China, [2] People's Hospital of Henan Province, Zhengzhou, China, [3] Institute of Psychology and Behavioral Science, Shanghai Jiao Tong University, Shanghai, China, [4] Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China

Despite the remarkable similarities between convolutional neural networks (CNN) and the human brain, CNNs still fall behind humans in many visual tasks, indicating that there still exist considerable differences between the two systems. Here, we leverage adversarial noise (AN) and adversarial interference (AI) images to quantify the consistency between neural representations and perceptual outcomes in the two systems. Humans can successfully recognize AI images as the same categories as their corresponding regular images but perceive AN images as meaningless noise. In contrast, CNNs can recognize AN images similar as corresponding regular images but classify AI images into wrong categories with surprisingly high confidence. We use functional magnetic resonance imaging to measure brain activity evoked by regular and adversarial images in the human brain, and compare it to the activity of artificial neurons in a prototypical CNN—AlexNet. In the human brain, we find that the representational similarity between regular and adversarial images largely echoes their perceptual similarity in all early visual areas. In AlexNet, however, the neural representations of adversarial images are inconsistent with network outputs in all intermediate processing layers, providing no neural foundations for the similarities at the perceptual level. Furthermore, we show that voxel-encoding models trained on regular images can successfully generalize to the neural responses to AI images but not AN images. These remarkable differences between the human brain and AlexNet in representation-perception association suggest that future CNNs should emulate both behavior and the internal neural presentations of the human brain.

Keywords: adversarial images, convolutional neural network, human visual cortex, functional magnetic resonance imaging, representational similarity analysis, forward encoding model

OPEN ACCESS

Edited by:
Tolga Cukur,
Bilkent University, Turkey

Reviewed by:
Shinji Nishimoto,
Osaka University, Japan
Mo Shahdloo,
University of Oxford, United Kingdom

***Correspondence:**
Ru-Yuan Zhang
ruyuanzhang@sjtu.edu.cn
Li Tong
tttocean_tl@hotmail.com

† These authors share senior authorship

Received: *08 March 2021*
Accepted: *28 June 2021*
Published: *05 August 2021*

Citation:
Zhang C, Duan X-H, Wang L-Y, Li Y-L, Yan B, Hu G-E, Zhang R-Y and Tong L (2021) Dissociable Neural Representations of Adversarially Perturbed Images in Convolutional Neural Networks and the Human Brain. Front. Neuroinform. 15:677925. doi: 10.3389/fninf.2021.677925

INTRODUCTION

The recent success of convolutional neural networks (CNNs) in many computer vision tasks inspire neuroscientists to consider them as a ubiquitous computational framework to understand biological vision (Jozwik et al., 2016; Yamins and DiCarlo, 2016). Indeed, a bulk of recent studies have demonstrated that visual features in CNNs can accurately predict many spatiotemporal

characteristics of brain activity (Agrawal et al., 2014; Yamins et al., 2014; Güçlü and van Gerven, 2015, 2017; Cichy et al., 2016; Hong et al., 2016; Horikawa and Kamitani, 2017; Khaligh-Razavi et al., 2017). These findings reinforce the view that modern CNNs and the human brain share many key structural and functional substrates (LeCun et al., 2015).

Despite the tremendous progress, current CNNs still fall short in several visual tasks. These disadvantages suggest that critical limitations still exist in modern CNNs (Grill-Spector and Malach, 2004). One potent example is adversarially perturbed images, a class of images that can successfully "fool" even the most state-of-the-art CNNs (Szegedy et al., 2013; Nguyen et al., 2015). Adversarial noise (AN) images (**Figure 1B**) look like meaningless noise to humans but can be classified by CNNs into familiar object categories with surprisingly high confidence (Nguyen et al., 2015). Adversarial interference (AI) images are generated by adding a small amount of special noise to regular images (**Figure 1C**). The special noise looks minimal to humans but severely impairs CNNs' recognition performance (Szegedy et al., 2013). *Perception* here can be operationally defined as the output labels of a CNN and object categories reported by humans. Therefore, adversarial images present a compelling example of double-dissociation between CNNs and the human brain, because artificially created images can selectively alter perception in one system without significantly impacting the other one.

It remains unclear the neural mechanisms underlying the drastically different visual behavior between CNNs and the human brain with respect to adversarial images. In particular, why do the two systems receive similar stimulus inputs but generate distinct perceptual outcomes? In the human brain, it has been known that the neural representations in low-level visual areas mostly reflect stimulus attributes whereas the neural representations in high-level visual areas mostly reflect perceptual outcomes (Grill-Spector and Malach, 2004; Wandell et al., 2007). For example, the neural representational similarity in human inferior temporal cortex is highly consistent with perceived object semantic similarity (Kriegeskorte et al., 2008). In other words, there exists a well-established representation-perception association in the human brain.

This processing hierarchy is also a key feature of modern CNNs. If the representational architecture in CNNs truly resembles the human brain, we should expect similar neural substrates supporting CNNs' "perception." For CNNs, AI images and regular images are more similar at the pixel level but yield different perceptual outcomes. By contrast, AN images and regular images are more similar at the "perceptual" level. We would expect that AI and regular images have more similar neural representations in low-level layers while AN and regular images have similar neural representations in high-level layers. In other words, there must exist at least one high-level representational layer that supports the same categorical perception of AN and regular images, similar to the representation-perception association in the human brain. However, as we will show later in this paper, we find no representational pattern that supports RE-AN perceptual similarity in all intermediate representation layers except the output layer.

The majority of prior studies focused on revealing similarities between CNNs and the human brain. In this paper, we instead leverage adversarial images to examine the differences between the two systems. We particularly emphasize that delineating the differences here does not mean to object CNNs as a useful computational framework for human vision. On the contrary, we acknowledge the promising utilities of CNNs in modeling biological vision but we believe it is more valuable to understand differences rather than similarities such that we are in a better position to eliminate these discrepancies and construct truly brain-like machines. In this study, we use a well-established CNN—AlexNet and investigate the activity of artificial neurons toward adversarial images and their corresponding regular images. We also use functional magnetic resonance imaging (fMRI) to measure the cortical responses evoked by RE and adversarial images in humans. Representational similarity analysis (RSA) and forward encoding modeling allow us to directly contrast representational geometries within and across systems to understand the capacity and limit of both systems.

MATERIALS AND METHODS

Ethics Statement

All experimental protocols were approved by the Ethics Committee of the Henan Provincial People's Hospital. All research was performed in accordance with relevant guidelines and regulations. Informed written consent was obtained from all participants.

Subjects

Three healthy volunteers (one female and two males, aged 22∼28 years) participated in the study. The subject S3 was the author C.Z. The other two subjects were naïve to the purpose of this study. All subjects were monolingual native-Chinese speakers and right-handed. All subjects had a normal or corrected-to-normal vision and considerable experience of fMRI experiments.

Convolutional Neural Network

We chose AlexNet and implemented it using the Caffe deep learning framework (Deng et al., 2009; Krizhevsky et al., 2012). AlexNet consists of five convolutional layers and three fully-connected layers (**Figure 1D**). The five convolutional layers each have 96, 256, 384, 384, and 256 linear convolutional kernels. The three fully-connected layers each have 4096, 4096, and 1000 artificial neurons. All convolutional layers perform linear convolution and rectified linear unit (ReLU) gating. Spatial max pooling is used only in layers 1, 2, and 5 to promote the spatial invariance of sensory inputs. In layers 1 and 2, local response normalization implements the inhibitory interactions across channels in a convolutional layer. In other words, the strong activity of a neuron in the normalization pool suppresses the activities of other neurons. Lateral inhibition of neurons is a well-established phenomenon in visual neuroscience and has proven to be critical to many forms of visual processing (Blakemore et al., 1970). The ReLU activation function and

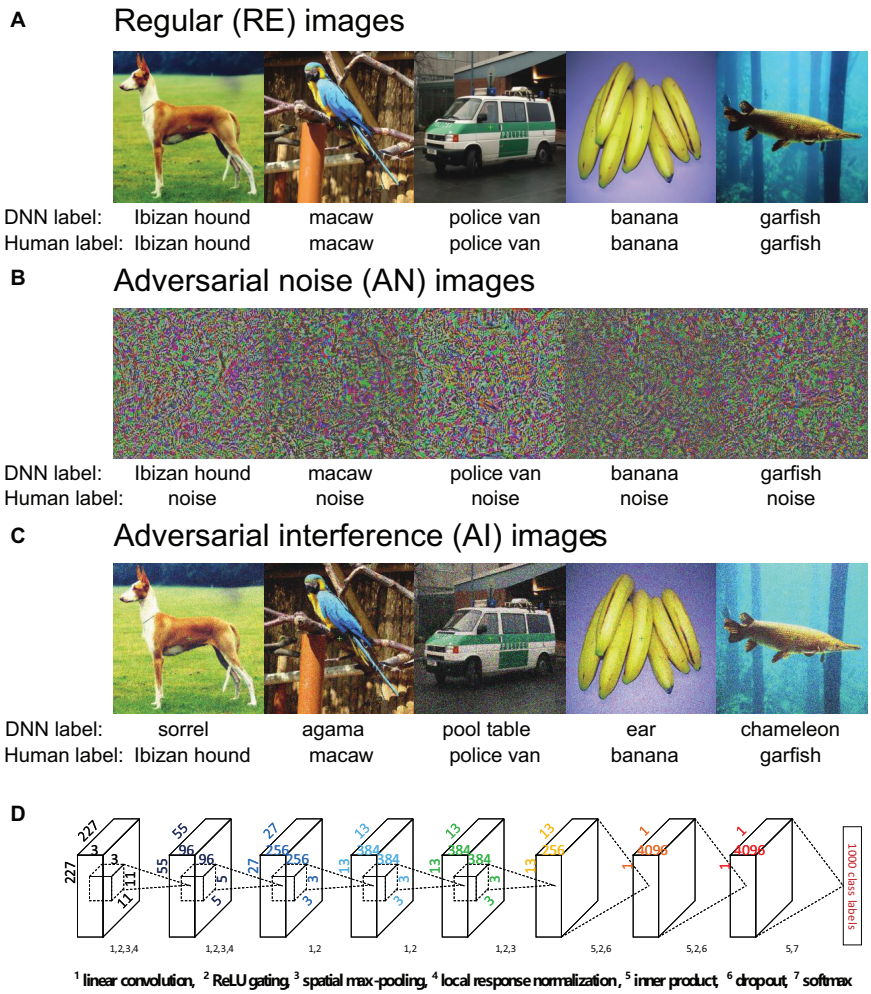


FIGURE 1 | (A–C) Example regular (RE, panel **A**), adversarial noise (AN, panel **B**) images and adversarial interference (AI, panel **C**) images. The five AN and five AI images one-by-one correspond to the five RE images. The labels provided by AlexNet and humans are listed under the images. The AI images contain a small amount of special image noise but overall look similar to the corresponding RE images. Humans can easily recognize the AI images as corresponding categories but the AN images as noise. AlexNet can classify the AN images into corresponding categories with over 99% confidence, but recognize the AI images as wrong categories. **(D)** The architecture of AlexNet. Details have been documented in Krizhevsky et al. (2012). Each layer uses some or all the following operations: linear convolution, ReLU gating, spatial max-pooling, local response normalization, inner product, dropout and softmax.

dropout are used in fully-connected layers 6 and 7. Layer 8 uses the softmax function to output the probabilities for 1000 target categories. In our study, all images were resized to 227 × 227 pixels in all three RGB color channels.

Image Stimuli

Regular Images

Regular (RE) images (**Figure 1A**) in our study were sampled from the ImageNet database (Deng et al., 2009). ImageNet is currently the most advanced benchmark database on which almost all state-of-the-art CNNs are trained for image classification. We selected one image (width and height > 227 pixels and aspect ratio > 2/3 and < 1.5) from each of 40 representative object categories. AlexNet can classify all images into their corresponding categories with probabilities greater than 0.99. The 40 images can be evenly divided into 5 classes: dogs, birds, cars, fruits, and aquatic animals (see **Supplementary Table 1** for details).

Adversarial Images

Adversarial images include adversarial noise (AN) (**Figure 1B**) and adversarial interference (AI) images (**Figure 1C**). A pair of AN and AI images were generated for each RE image. As such, a total of 120 images (40 RE + 40 AN + 40 AI) were used in the entire experiment.

The method to generate AN images has been documented in Nguyen A et al. (Nguyen et al., 2015). We briefly summarize the method here. We first used the averaged image of all images in ImageNet as the initial AN image. Note that the category label of the corresponding RE image was known, and AlexNet had been fully trained. As such, we first fed the initial AN image to

AlexNet and forwardly computed the probability for the correct category. This probability was expected to be initially low. We then used the backpropagation method to transduce error signals from the top layer to image pixel space. Pixel values in the initial AN image were then adjusted accordingly to enhance the classification probability. This process of forwarding calculation and backpropagation was iterated many times until the pixel values of AN image converged.

We also included an additional regularization item to control the overall intensity of the image. Formally, let $P_c(I)$ be the probability of class c (RE image label) given an image I. We would like to find an L_2-regularized image I^*, such that it maximizes the following objective:

$$I^* = \arg\max_I \; P_c(I) - \lambda \, ||I\text{-}I_{mean}||_2^2 \,, \qquad (1)$$

where, λ is the regularization parameter and I_{mean} is the grand average of all images in ImageNet. Finally, all the probabilities of generated AN images used in our experiment being classified into RE images were greater than 0.99. Note that the internal structure (i.e., all connection weights) of AlexNet was fixed throughout the entire training process, and we only adjusted pixel values in input AN images.

The AI images were generated by adding noise to the RE images. For an RE image (e.g., dog), a wrong class label (e.g., bird) was pre-selected (see **Supplementary Table 1** for details). We then added random noise (uniform distribution $-5\sim5$) to every pixel in the RE image. The resulted image was kept if the probability of this image being classified into the wrong class (i.e., bird) increased, and was discarded otherwise. This procedure was repeated many times until the probability for the wrong class exceeded 0.5 (i.e., wrong class label as the top1 label). We deliberately choose 0.5 because under this criteria the resulted images were still visually comparable to the RE images. A higher stopping criteria (e.g., 0.99) may overly load noises and substantially reduce image visibility. We further used the similar approach as AN images (change the I_{mean} in Eq. 1 to I_{RE}.) to generate another set of AI images (with a probability of over 0.99 to be classified into the "wrong" class) and confirmed that the results in AlexNet RSA analyses did not substantially change under this regime (see **Supplementary Figure 4**). We adopted the former not the latter approach in our fMRI experiment because the differences between the AI and the RE images were so small that the human eye can hardly see it in the experiment. This is meaningless for an fMRI experiment as the AI and the RE images look "exactly" the same, which is equivalent to present the identical images twice.

Apparatus

All computer-controlled stimuli were programmed in Eprime 2.0 and presented using a Sinorad LCD projector (resolution 1920 × 1080 at 120 Hz; size 89 cm × 50 cm; viewing distance 168 cm). Stimuli were projected onto a rear-projection monitor located over the head. Subjects viewed the monitor via a mirror mounted on the head coil. Behavioral responses were recorded by a button box.

fMRI Experiments

Main Experiment

Each subject underwent two scanning sessions in the main experiment. In each session, half of all images (20 images x 3 RE/AN/AI = 60 images) were presented. Each session consisted of five scanning runs, and each run contained 129 trials (2 trials per image and 9 blank trials). The image presentation order was randomized within a run. In a trial, a blank lasted 2 s and was followed by an image (12° × 12°) of 2 s. A 20 s blank period was included to the beginning and the end of each run to establish a good baseline and compensate for the initial insatiability of the magnetic field. A fixation point (0.2° × 0.2°) was shown at center-of-gaze, and participants were instructed to maintain steady fixation throughout a run. Participants pressed buttons to perform an animal judgment task—whether an image belongs to animals. The task aimed to engage subjects' attention onto the stimuli.

Retinotopic Mapping and Functional Localizer Experiments

A retinotopic mapping experiment was also performed to define early visual areas, as well as two functional localizer experiments to define lateral occipital (LO) lobe and human middle temporal lobe (hMT+).

The retinotopic experiment used standard phase-encoding methods (Engel et al., 1994). Rotating wedges and expanding rings were filled by textures of objects, faces, and words, and were presented on top of achromatic pink-noise backgrounds (http://kendrickkay.net/analyzePRF/). Early visual areas (V1–V4) were defined on the spherical cortical surfaces of individual subjects.

The two localizer experiments were used to create a more precise LO mask (see region-of-interest definition section below). Each localizer experiment contained two runs. In the LO localizer experiment, each run consisted of 16 stimulus blocks and 5 blank blocks. Each run began with a blank block, and a blank block appeared after every 4 stimulus blocks. Each block lasted 16 s. Intact images and their corresponding scrambled images were alternately presented in a stimulus block. Each stimulus block contained 40 images (i.e., 20 intact + 20 scramble images). Each image (12° × 12°) lasted 0.3 s and was followed by a 0.5 s blank.

In the hMT+ localizer experiment, each run contained 10 stimulus blocks, and each block lasted 32 s. In a block, a static dot stimulus (24 s) and a moving-dot stimulus (8 s) were alternately presented. All motion stimuli subtended a 12° × 12° square area on a black background. An 8 s blank was added to the beginning and the end of each run. Note that hMT+ here is only used to remove motion-selective vertices from the LO mask (see Region-Of-Interest definitions). We did not analyze motion signals in hMT+ as all our images were static.

MRI Data Acquisition

All MRI data were collected using a 3.0-Tesla Siemens MAGNETOM Prisma scanner and a 32-channel head coil at the Department of Radiology at the People's Hospital of Henan Province.

An interleaved T2*-weighted, single-shot, gradient-echo echo-planar imaging (EPI) sequence was used to acquire

functional data (60 slices, slice thickness 2 mm, slice gap 0 mm, field of view 192 × 192 mm^2, phase-encode direction anterior-posterior, matrix size 96 × 96, *TR/TE* 2000/29 ms, flip angle 76°, nominal spatial resolution 2 × 2 × 2 mm^3). Three B0 fieldmaps were acquired to aid post-hoc correction for EPI spatial distortion in each session (resolution 2 × 2 × 2 mm^3, TE_1 4.92 ms, TE_2 7.38 ms, *TA* 2.2 min). In addition, high-resolution T1-weighted anatomical images were also acquired using a 3D-MPRAGE sequence (*TR* 2300 ms, *TE* 2.26 ms, *TI* 900 ms, flip angle 8°, field of view 256 × 256 mm^2, voxel size 1. × 1. × 1. mm^3).

MRI Data Preprocessing

The pial and the white surfaces of subjects were constructed from T1 volume using FreeSurfer software (http://surfer.nmr.mgh.harvard.edu). An intermediate gray matter surface between the pial and the white surfaces was also created for each subject.

Our approach for dealing with EPI distortion followed Kay et al. (2019). Fieldmaps acquired in each session were phase-unwrapped using the FSL utility prelude (version 2.0) with flags -s -t 0. We then regularized the fieldmaps by performing 3D local linear regression using an Epanechnikov kernel with radius 5 mm. We used values in the magnitude component of the fieldmap as weights in the regression in order to improve robustness of the field estimates. This regularization procedure removes noise from the fieldmaps and imposes spatial smoothness. Finally, we linearly interpolated the fieldmaps over time, producing an estimate of the field strength for each functional volume acquired.

For functional data, we discarded the data points of the first 18 s in the main experiment, the first 14 s in the LO localizer experiment, and the first 6 s in the hMT+ localizer experiment. This procedure ensures a 2 s blank was kept before the first task block in all three experiments.

The functional data were initially volume-based pre-processed by performing one temporal and one spatial resampling. The temporal resampling realized slice time correction by executing one cubic interpolation for each voxel's time series. The spatial resampling was performed for EPI distortion and head motion correction. The regularized time-interpolated field maps were used to correct EPI spatial distortion. Rigid-body motion parameters were then estimated from the undistorted EPI volumes with the SPM5 utility spm_realign (using the first EPI volume as the reference). Finally, the spatial resampling was achieved by one cubic interpolation on each slice-time-corrected volume (the transformation for correcting distortion and the transformation for correcting motion are concatenated such that a single interpolation is performed).

We co-registered the average of the pre-processed functional volumes obtained in a scan session to the T1 volume (rigid-body transformation). In the estimation of the co-registration alignment, we used a manually defined 3D ellipse to focus the cost metric on brain regions that are unaffected by gross susceptibility effects (e.g., near the ear canals). The final result of the co-registration is a transformation that indicates how to map the EPI data to the subject-native anatomy.

With the anatomical co-registration complete, the functional data were re-analyzed using surface-based pre-processing. The reason for this two-stage approach is that the volume-based pre-processing is necessary to generate the high-quality undistorted functional volume that is used to determine the registration of the functional data to the anatomical data. It is only after this registration is obtained that the surface-based pre-processing can proceed.

In surface-based pre-processing, the exact same procedures associated with volume-based pre-processing are performed, except that the final spatial interpolation is performed at the locations of the vertices of the intermediate gray matter surfaces. Thus, the only difference between volume- and surface-based pre-processing is that the data are prepared either on a regular 3D grid (volume) or an irregular manifold of densely spaced vertices (surface). The entire surface-based pre-processing ultimately reduces to a single temporal resampling (to deal with slice acquisition times) and a single spatial resampling (to deal with EPI distortion, head motion, and registration to anatomy). Performing just two simple pre-processing operations has the benefit of avoiding unnecessary interpolation and maximally preserving spatial resolution (Kang et al., 2007; Kay and Yeatman, 2017; Kay et al., 2019). After this pre-processing, time-series data for each vertex of the cortical surfaces were ultimately produced.

General Linear Modeling

We estimated the vertex responses (i.e., beta estimates from GLM modeling) of all stimulus trials in the main experiment using the GLMdenoise method (Kay et al., 2013). All blank trials were modeled as a single predictor. This analysis yielded beta estimations of 241 conditions (120 images × 2 trials + 1 blank trial). Notably, we treated two presentations of the same image as two distinct predictors in order to calculate the consistency of the response patterns across the two trials.

Region-of-Interest Definitions

Based on the retinotopic experiment, we calculated the population receptive field (pRF) (http://kendrickkay.net/analyzePRF) of each vertex and defined low-level visual areas (V1–V4) based on the pRF maps. To define LO, we first selected vertices that show significantly higher responses to intact images than scrambled images (two-tails *t*-test, $p < 0.05$, uncorrected). In addition, hMT+ was defined as the area that shows significantly higher responses to moving than static dots (two-tails *t*-test, $P < 0.05$, uncorrected). The intersection vertices between LO and hMT+ were then removed from LO.

Vertex Selection

To further select task-related vertices in each ROI (**Figure 2A**), we performed a searchlight analysis on flattened 2D cortical surfaces (Chen et al., 2011). For each vertex, we defined a 2D searchlight disk with 3 mm radius. The geodesic distance between two vertices was approximated by the length of the shortest path between them on the flattened surface. Given the vertices in the disk, we calculated the representational dissimilarity matrices (RDM) of all RE images for each of the two presentation trials. The two RDMs were then compared (Spearman's R) to show the consistency of activity patterns across the two trials. Here rank-correlation (e.g., Spearman's R) is used as it was recommended

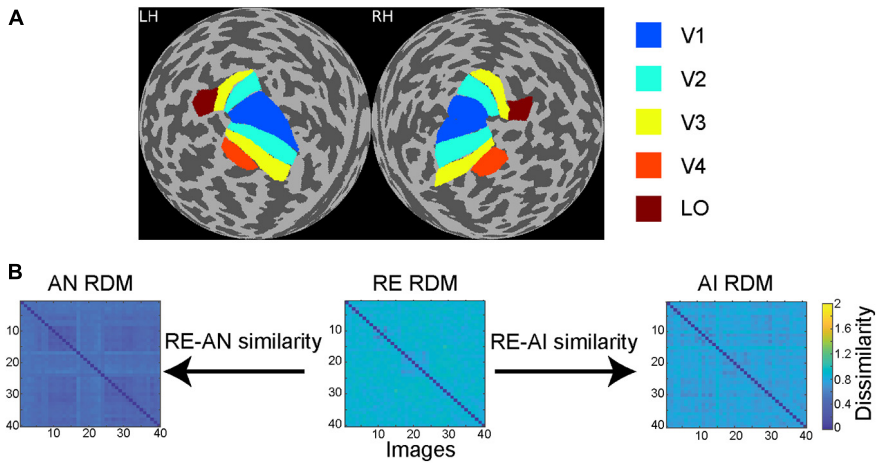


FIGURE 2 | (A) Regions of interest (ROIs) in a sample subject. Through retinotopic mapping and functional localizer experiments, we identified five ROIs—V1, V2, V3, V4 and lateral occipital (LO) cortex—in both left (LH) and right (RH) hemispheres. **(B)** Calculation of RE-AN and RE-AI similarity. For each CNN layer or brain ROI, three RDMs are calculated with respect to the three types of images. We then calculate the Spearman correlation between the AN and the RE RDMs, obtaining the RE-AN similarity. Similarly, we can calculate the RE-AI similarity.

when comparing two RDMs (Kriegeskorte et al., 2008; Nili et al., 2014).

The 200 vertices (100 vertices from each hemisphere) with the highest correlation values were selected in each ROI for further analysis (**Figure 3**). Note that vertex selection was only based on the responses to the RE images and did not involve any response data for the AN and the AI images. We also selected a total of 400 vertices in each area and we found our results held. The results are shown in **Supplementary Figure 2**.

Representational Similarity Analysis

We applied RSA separately to the activity in the CNN and the brain.

RSA on CNN Layers and Brain ROIs

For one CNN layer, we computed the representational dissimilarity between every pair of the RE images, yielding a 40 × 40 RDM (i.e., RDM_{RE}) for the RE images. Similarly, we obtained the other two RDMs each for the AN (i.e., RDM_{AN}) and the AI images (i.e., RDM_{AI}). We then calculated the similarity between the three RDMs as follows:

$$R_{RE-AN} = corr(RDM_{RE}, RDM_{AN}), \quad (2)$$

$$R_{RE-AI} = corr(RDM_{RE}, RDM_{AI}), \quad (3)$$

This calculation generated one RE-AN similarity value and one RE-AI similarity value for that CNN layer (see **Figure 2B**). We repeated the same analysis above on the human brain except that we used the activity of vertices in a brain ROI.

In a given ROI or AlexNet layer, we first resampled 80% voxels or artificial neurons without replacement (**Supplementary Figure 5**). In each sample, we calculated RE, AI, and AN RDM, and calculated the difference between RE-AI similarity and RE-AN similarity, obtaining one difference value. This was done 1000 times, yielding 1000 different values as the baseline distribution for RE-AI and RE-AN difference. This method is used for examining the relative difference between the RE-AN and the RE-AN similarities.

To construct the null hypotheses for the absolute RE-AN and the RE-AI similarities, in each voxel or artificial neuron sample, we further permuted the image labels with respect to their corresponding activities for the RE images (**Supplementary Figure 6**). In other words, an image label may be paired with a wrong activity pattern. We then recalculated the RE-AN and the RE-AI similarities. In this way, 1000 RE-AN and 1000 RE-AI similarity values were generated. The two distributions consisting of 1000 values were regarded as the null hypothesis distributions of RE-AN or RE-AI, respectively.

In addition, the Mann-Kendall test was applied to assess the monotonic upward or downward trend of the RE-AN similarities over CNN layers. The Mann-Kendall test can be used in place of a parametric linear regression analysis, which can be used to test if the slope of the estimated linear regression line is different from zero.

In order to verify the statistical effectiveness of the fMRI experimental results of the three subjects, we used the G*Power tool (Faul et al., 2009) to re-analyze our experimental results. For each ROI, we carried out a paired *t*-test (i.e., "means: difference between two dependent means (matched pairs)" in G*Power) on the RE-AI similarities and the RE-AN similarities of the three subjects. We calculated three RE-AI/RE-AN difference values (i.e., the height difference between blue and red bars in **Figure 3**), each for one subject. The effect size was determined from the mean and SD of the difference values. We first set the type of power analysis to "*post hoc*: compute achieved power – given α, sample size, and effect size" to estimate the statistical power given $N = 3$. The statistical power (1-β error probability, α error probability was set to 0.05) was then calculated. We then set the type of power analysis to "*a priori*: compute required sample

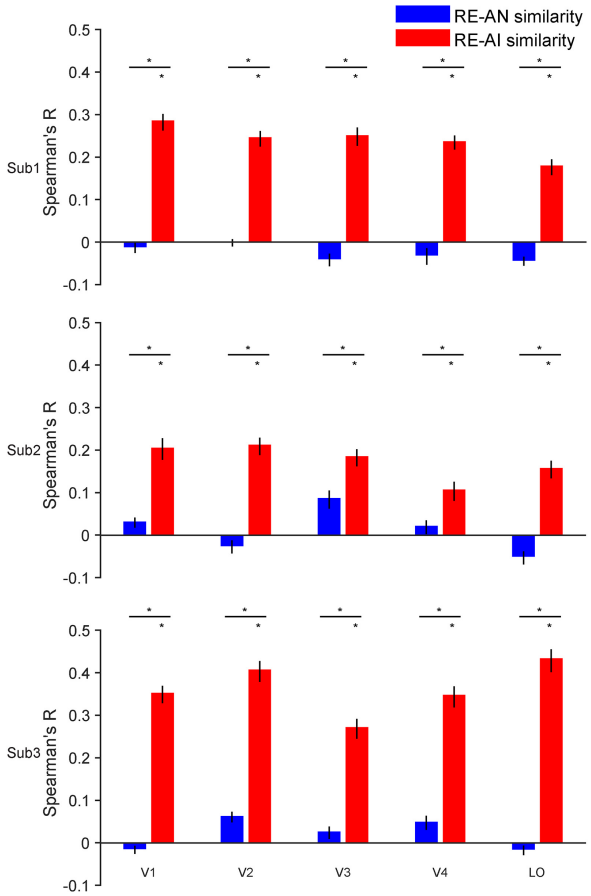


FIGURE 3 | RE-AI and RE-AN similarities in the human brain. Three subplots indicate the three human subjects. In all five brain ROIs, the RE-AI (red bars) similarities are substantially higher than the RE-AN (blue bars) similarities. Error bars are 95% confidence intervals of similarity values by bootstrapping vertices in one brain ROI (see Methods). The black asterisks above bars indicate that the similarity values are significantly different from null hypotheses (permutation test, $p < 0.05$, see Methods).

size – given α, power, and effect size," and calculated the estimated minimum required sample size to achieve a statistical power of 0.8 with the current statistics.

Searchlight RSA

We also performed a surface-based searchlight analysis in order to show the cortical topology of the RE-AN and the RE-AI similarity values. For each vertex, the same 2D searchlight disk was defined as above. We then repeated the same RSA on the brain, producing two cortical maps with respect to the RE-AN and RE-AI similarity values.

Forward Encoding Modeling

Here, forward encoding models assume that the activity of a voxel in the brain can be modeled as the linear combination of the activity of artificial neurons in CNNs. Thus, forward encoding modeling can bridge the representations of the two systems. Thus, forward encoding modeling can bridge the representations of the two systems. This is also the typical approach in existing related works (Güçlü and van Gerven, 2015; Kell et al., 2018).

We first trained the forward encoding models only based on the RE images data in the brain and the CNN. For the response sequence $y = \{y_1, \ldots, y_d\}^{\mathrm{T}}$ of one vertex to the 40 RE images, it is expressed as Eq. (4):

$$\mathrm{y} = \mathrm{Xw}, \tag{4}$$

X is an m-by-(n+1) matrix, where m is the number of training images (i.e., 40), and n is the number of units in one CNN layer. The last column of **X** is a constant vector with all elements equal to 1. **w** is an (n+1)-by−1 unknown weighting matrix to solve. Because the number of training samples m was less than the number of units n in all CNN layers, we imposed an additional sparse constraint on the forward encoding models to avoid overfitting:

$$\min_{\mathrm{w}} ||\mathrm{w}||_0 \quad \text{subject to} \quad \mathrm{y} = \mathrm{Xw}, \tag{5}$$

Sparse coding has been widely suggested and used in both neuroscience and computer vision (Vinje and Gallant, 2000; Cox and Savoy, 2003). We used the regularized orthogonal matching pursuit (ROMP) method to solve the sparse representation problem. ROMP is a greedy method developed by Needell D and R Vershynin (Needell and Vershynin, 2009) for sparse recovery. Features for prediction can be automatically selected to avoid overfitting. For the selected 200 vertices in each human ROI, we established 8 forward encoding models corresponding to the 8 CNN layers. This approach yielded a total of 40 forward encoding models (5 ROIs × 8 layers) for one subject.

Based on the train forward encoding models, we calculated the Pearson correlation between the empirically measured and model-predicted response patterns evoked by the adversarial images. To test the prediction accuracy against null hypotheses, we randomized the image labels and performed permutation tests as described above. Specifically, we resampled 80% vertices in a brain ROI 1000 times without replacement and in each sample recalculated the mean response prediction accuracy, resulting in a bootstrapped accuracy distribution with 1000 mean response prediction accuracy values (**Supplementary Figure 7**). The upper and lower bounds of the 95% confidence intervals were derived from the bootstrapped distribution. Similarly, we compared the bootstrapped distributions of two types of adversarial images to derive the statistical difference between the RE-AI and the RE-AN similarity.

RESULTS

Dissociable Neural Representations of Adversarial Images in AlexNet and the Human Brain

Human Brain

For one brain ROI, we calculated the representational dissimilarity matrix (i.e., 40 x 40 RDM) for each of the

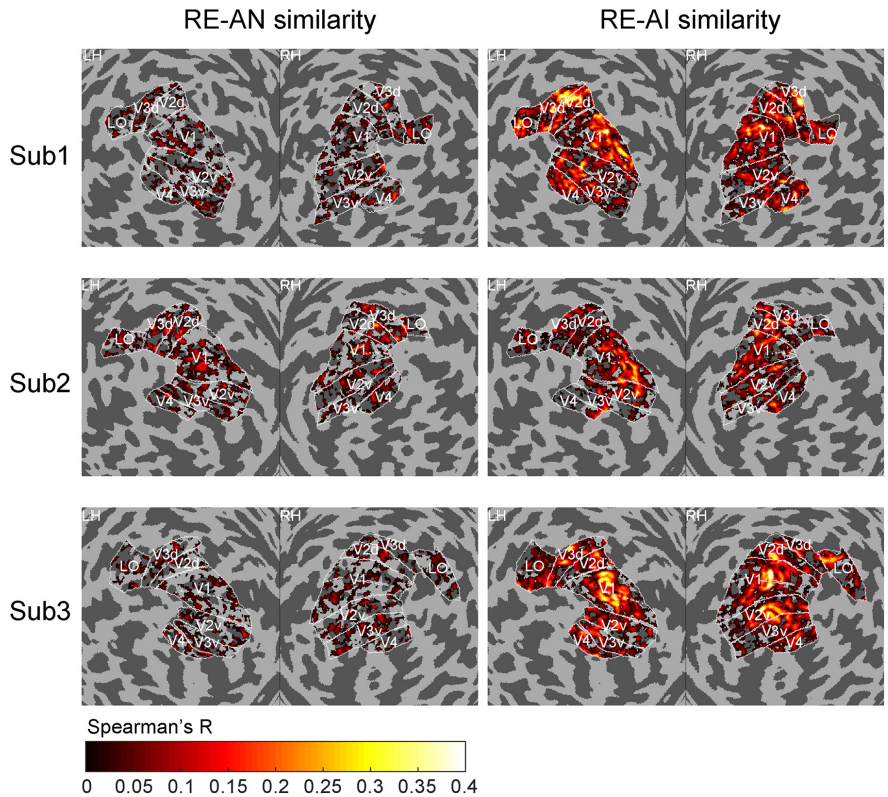


FIGURE 4 | Cortical topology of RE-AI and RE-AN similarities. The RE-AI similarities are overall higher than the RE-AN similarities across all early visual areas in the human brain.

three image types. We then calculated the RE-AN similarity—the correlation between the RDM of the RE images and that of the AN images, and the RE-AI similarity between the RE images and the AI images.

We made three major observations. First, the RE-AI similarities were significantly higher than null hypotheses in almost all ROIs in the three subjects (red bars in **Figure 3**, permutation test, all p-values < 0.005, see Methods for the deviation of null hypotheses). Conversely, this was not true for the RE-AN similarities (blue bar in **Figure 3**, permutation test, only four p-values < 0.05 in 3 subjects x 5 ROI = 15 tests). Third and more importantly, we found significantly higher RE-AI similarities than the RE-AN similarities in all ROIs (**Figure 3**, bootstrap test, all p-values < 0.0001). These results suggest that the neural representations of the AI images, compared with the AN images, are much more similar to that of the corresponding RE images. Notably, this representational structure is also consistent with the perceptual similarity of the three types of images in humans. In other words, the neural representations of all images in the human brain largely echo their perceptual similarity.

In addition, the results of the statistical power analysis showed that the final average power (1-β error probability, α error probability was set to 0.05, $N = 3$) across five ROIs for the paired t-test on RE-AI similarities and RE-AN similarities of the three subjects equaled 0.818 (V1:0.911, V2: 0.998, V3:0.744, V4:0.673, LO:0.764). And the average minimum required sample size was 2.921 (V1:2.623, V2:2.131, V3:3.209, V4:3.514, LO:3.129, the power was set to 0.8). In other words, the number of subjects can meet the minimum statistical power.

We also performed a searchlight analysis to examine the cortical topology of the neural representations. The searchlight analysis used the same calculation as above (see Methods). We replicated the results (see **Figure 4**) and found a distributed pattern of higher RE-AI similarities in the early human visual cortex. In addition, we expanded our searchlight analysis for broader regions (see **Supplementary Figure 3**) and obtained the qualitatively same main results.

AlexNet

We repeated our analyses above in AlexNet and again made three observations. First, the RE-AI similarities were higher than null hypotheses across all layers (**Figure 5**, permutation test, all p-values < 0.001), and the RE-AI similarities declined from low to high layers (Mann–Kendall test, $p = 0.009$). Second, the RE-AN similarities were initially low (p-values > 0.05 in layers 1–2) but then dramatically increased (Mann–Kendall test, $p < 0.001$) and became higher than the null hypotheses from layer 3 (all p-values < 0.05 in layers 3–8). Third and most importantly, we found that the RE-AN similarities were not higher than the RE-AI similarities in all intermediate layers (i.e., layers 1–7, bootstrap

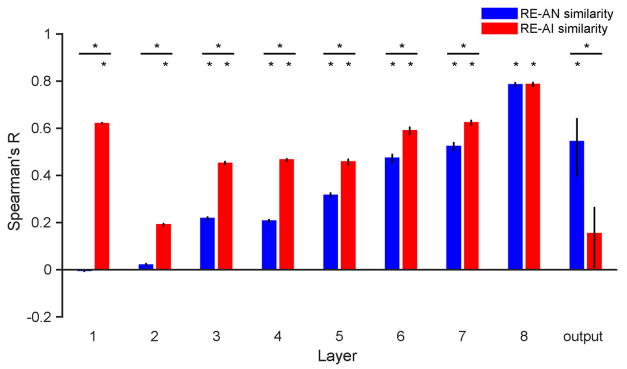


FIGURE 5 | RE-AN and RE-AI similarities across layers in AlexNet. the RE-AN similarity increases and the RE-AI similarities decline along the processing hierarchy. The RE-AN similarities are not higher than the RE-AI similarities in all representational layers (i.e., layer 1–7). Error bars indicate 95% bootstrapped confidence intervals (see Methods).

test, all p-values < 0.05, layer 7, $p = 0.375$) except the output layer (i.e., layer 8, $p < 0.05$).

These results are surprising because it suggests that neural representations of the AI images, compared with the AN images, are more similar to the representations of the RE images. However, the output labels of the AN images are similar to those of the corresponding RE images in AlexNet. In other words, there exists substantial inconsistency between the representational similarity and perceptual similarity in AlexNet. We emphasize that, assuming that in order for two images look similar, there must be at least some neural populations somewhere in a visual system that represents them similarly. But, astonishingly, we found no perception-compatible neural representations in any representational layer. Also, the transformation from layer 7 to the output layer is critical and eventually renders the RE-AN similarity higher than the RE-AI similarity in the output layer. This is idiosyncratic because AlexNet does not implement effective neural codes of objects in representational layers beforehand but the last transformation reverses the relative RDM similarity of the three types of images. This is drastically different from the human brain that forms correct neural codes in all early visual areas.

Forward Encoding Modeling Bridges Responses in AlexNet and Human Visual Cortex

The RSA above mainly focuses on the comparisons across image types within one visual system. We next used forward encoding modeling to directly bridge neural representations across the two systems. Forward encoding models assume that the activity of a voxel in the brain can be modeled as the linear combination of the activity of multiple artificial neurons in CNNs. Following this approach, we trained a total of 40 (5 ROIs x 8 layers) forward encoding models for one subject using regular images. We then tested how well these trained forward encoding models can generalize to the corresponding adversarial images. The rationale is that, if the brain and AlexNet process images in a similar fashion, the forward encoding models trained on the RE images should transfer to the adversarial images, and vice versa if not.

We made two major findings here. First, almost all trained encoding models successfully generalized to the AI images (**Figure 6**, warm color bars, permutation test, p-values < 0.05 for 113 out of the 120 models for three subjects) but not to the AN images (**Figure 6**, cold color bars, permutation test, p-values > 0.05 for 111 out of the 120 models). Second, the forward encoding models exhibited much stronger predictive power on the AI images than the AN images (bootstrap test, all p-values < 0.05, except the encoding model based on layer 8 for LO in subject 2, $p = 0.11$). These results suggest that the functional correspondence between AlexNet and the human brain only holds when processing RE and AI images but not AN images. This result is also consonant with the RSA above and demonstrates that both systems treat RE and AI images similarly, but AN images very differently. But again, note that AlexNet exhibits the opposite behavioral pattern of human vision.

DISCUSSION AND CONCLUSION

Given that current CNNs still fall short in many tasks, we use adversarial images to probe the functional differences between a prototypical CNN—AlexNet, and the human visual system. We make three major findings. First, the representations of AI images, compared with AN images, are more similar to the representations of corresponding RE images. These representational patterns in the brain are consistent with human percepts (i.e., perceptual similarity). Second, we discover a representation-perception disassociation in all intermediate layers in AlexNet. Third, we use forward encoding modeling to link neural activity in both systems. Results show that the processing of RE and AI images are quite similar but both are significantly different from AN images. Overall, these observations demonstrate the capacity and limit of the similarities between current CNNs and human vision.

Abnormal Neural Representations of Adversarial Images in CNNs

To what extent neural representations reflect physical or perceived properties of stimuli is a key question in modern vision science. In the human brain, researchers have found that early visual processing mainly processes low-level physical properties of stimuli, and late visual processing mainly supports high-level categorical perception (Grill-Spector and Malach, 2004). We ask a similar question here—to what extent neural representations in CNNs or the human brain reflect their conscious perception.

One might argue that the representation-perception disassociation in AlexNet is trivial, given that we already know that AlexNet exhibits opposite behavioral patterns compared to human vision. But we believe thorough quantifications of their neural representations in both systems are still of great value. First, neural representations do not necessarily follow our conscious perception, and numerous neuroscience studies have shown disassociated neural activity and perception

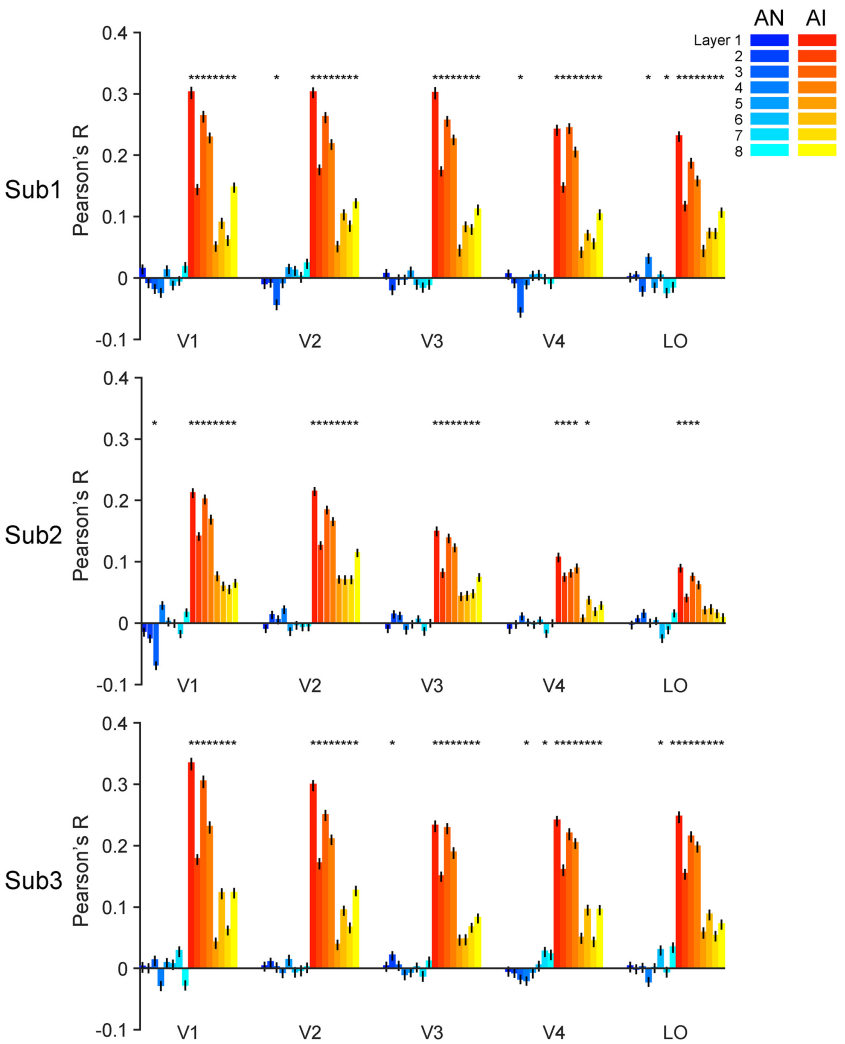


FIGURE 6 | Accuracy of forward encoding models trained on RE images and then tested on adversarial images. After the models are fully trained on the RE images, we input the adversarial images as inputs to the models can predict corresponding brain responses. The *y*-axis indicates the Pearson correlation between the brain responses predicted by the models and the real brain responses. The generalizability of forward encoding models indicates the processing similarity between the RE and AN (cool colors) or AI (warm colors) images. Error bars indicate 95% bootstrapped confidence intervals (see Methods).

in both the primate or human brain in many cases, such as visual illusion, binocular rivalry, visual masking (Serre, 2019). The question of representation-perception association lies at the center of the neuroscience of consciousness and should also be explicitly addressed in AI research. Second, whether representation and perception are consistent or not highly depends on processing hierarchy, which again needs to be carefully quantified across visual areas in the human brain and layers in CNNs. Here, we found no similar representations of AN and regular images in any intermediate layer in AlexNet even though they "look" similar. This is analogous to the scenario that we cannot decode any similar representational patterns of two images throughout a subject's brain, although the subject behaviorally reports the two images are similar.

Adversarial Images as a Tool to Probe Functional Differences Between the CNN and Human Vision

In computer vision, adversarial images impose problems on the real-life applications of artificial systems (i.e., adversarial attack) (Yuan et al., 2017). Several theories have been proposed to explain the phenomenon of adversarial images (Akhtar and Mian, 2018). For example, one possible explanation is that CNNs are forced to behave linearly in high dimensional spaces, rendering them vulnerable to adversarial attacks (Goodfellow et al., 2014b). Besides, flatness (Fawzi et al., 2016) and large local curvature of the decision boundaries (Moosavi-Dezfooli et al., 2017), as well as low flexibility of the networks (Fawzi et al., 2018) are all possible reasons. (Szegedy et al., 2013) has suggested that current CNNs

are essentially complex nonlinear classifiers, and this discriminative modeling approach does not consider generative distributions of data. We will further address this issue in the next section.

In this study, we focused on one particular utility of adversarial images—to test the dissimilarities between CNNs and the human brain. Note that although the effects of adversarial images indicate the deficiencies of current CNNs, we do not object to the approach to use CNNs as a reference to understand the mechanisms of the brain. Our study here fits the broad interests in comparing CNNs and the human brain in various aspects. We differ from other studies just because we focus on their differences. We do acknowledge that it is quite valuable to demonstrate functional similarities between the two systems. But we believe that revealing their differences, as an alternative approach, might further foster our understandings of how to improve the design of CNNs. This is similar to the logic of using ideal observer analysis in vision science. Although we know human visual behavior is not optimal in many situations, the comparison to an ideal observer is still meaningful as it can reveal some critical mechanisms of human visual processing. Also, we want to emphasize that mimicking the human brain is not the only way or even may not be the best way to improve CNN performance. Here, we only suggest a potential route given that current CNNs still fall short in many visual tasks as compared to humans.

Some recent efforts have been devoted to addressing CNN-human differences. For example, Rajalingham et al. (2018) found that CNNs explain human (or non-human primate) rapid object recognition behavior at the level of category but not individual images. CNNs better explain the ventral stream than the dorsal stream (Wen et al., 2017). To further examine their differences, people have created some unnatural stimuli/tasks, and our work on adversarial images follows this line of research. The rationale is that, if CNNs are similar to humans, they should exhibit the same capability in both ordinary and unnatural circumstances. A few studies adopted some other manipulations (Flesch et al., 2018; Rajalingham et al., 2018), such as manipulation of image noise (Geirhos et al., 2018) and distortion (Dodge and Karam, 2017).

Possible Caveats of CNNs in the Processing of Adversarial Images

Why CNNs and human vision behave differently on adversarial images, especially on AN images? We want to highlight three reasons and discuss the potential route to circumvent them.

First, current CNNs are trained to match the classification labels generated by humans. This approach is a discriminative modeling approach that characterizes the probability of $p(\text{class} \mid \text{image})$. Note that natural images only occupy a low-dimensional manifold in the entire image space. Under this framework, there must exist a set of artificial images in the image space that fulfills a classifier but does not belong to any distribution of real images. Humans cannot recognize AN images because humans do not merely rely on discriminative classifiers but instead perform Bayesian inference and take into consideration both likelihood $p(\text{image} \mid \text{class})$ and prior experience $p(\text{class})$. One approach to overcome this is to build generative deep models to learn latent distributions of images, such as variational autoencoders (Kingma and Welling, 2013) and generative adversarial networks (Goodfellow et al., 2014a).

Another advantage of deep generative models is to explicitly model the uncertainty in sensory processing and decision. It has been well-established in cognitive neuroscience that the human brain computes not only form a categorical perceptual decision, but also a full posterior distribution over all possible hidden causes given a visual input (Knill and Pouget, 2004; Wandell et al., 2007; Pouget et al., 2013). This posterior distribution is also propagated to downstream decision units and influences other aspects of behavior.

Third, more recurrent and feedback connections are needed. Numerous studies have shown the critical role of top-down processing in a wide range of visual tasks, including recognition (Bar, 2003; Ullman et al., 2016), tracking (Cavanagh and Alvarez, 2005), as well as other cognitive domains, such as memory (Zanto et al., 2011), language comprehension (Zekveld et al., 2006) and decision making (Fenske et al., 2006; Rahnev, 2017). In our results, the responses in the human visual cortex likely reflect the combination of feedforward and feedback effects whereas the activity in most CNNs only reflects feedforward inputs from earlier layers. A recent study has shown that recurrence is necessary to predict neural dynamics in the human brain using CNN features (Engel et al., 1994).

CONCLUDING REMARKS

In the present study, we compared neural representations of adversarial images in AlexNet and the human visual system. Using RSA and forward encoding modeling, we found that the neural representations of RE and AI images are similar in both systems but AN images were idiosyncratically processed in AlexNet. These findings open a new avenue to help design CNN architectures to achieve brain-like computation.

DISCLOSURE STATEMENT

All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (Henan Provincial People's Hospital) and with the Helsinki Declaration of 1975, as revised in 2008 (5). Informed consent was obtained from all patients for being included in the study.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Henan Provincial People's Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CZ, R-YZ, LT, and BY designed the research. CZ, X-HD, L-YW, G-EH, and LT collected the data. CZ and R-YZ analyzed the data and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Key Research and Development Plan of China under Grant 2017YFB1002502.

ACKNOWLEDGMENTS

We thank Pinglei Bao, Feitong Yang, Baolin Liu, and Huafu Chen for their invaluable comments on the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2021.677925/full#supplementary-material

REFERENCES

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J. L. (2014). Pixels to voxels: modeling visual representation in the human brain. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1407.5104 (accessed February 23, 2021).

Akhtar, N., and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: a survey. *IEEE Access.* 6, 14410–14430. doi: 10.1109/access.2018.2807385

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *J. Cogn. Neurosci.* 15, 600–609. doi: 10.1162/089892903321662976

Blakemore, C., Carpenter, R. H., and Georgeson, M. A. (1970). Lateral inhibition between orientation detectors in the human visual system. *Nature* 228, 37–39. doi: 10.1038/228037a0

Cavanagh, P., and Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends Cogn. Sci.* 9, 349–354. doi: 10.1016/j.tics.2005.05.009

Chen, Y., Namburi, P., Elliott, L. T., Heinzle, J., Soon, C. S., Chee, M. W., et al. (2011). Cortical surface-based searchlight decoding. *Neuroimage* 56, 582–592. doi: 10.1016/j.neuroimage.2010.07.035

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6:27755.

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/s1053-8119(03)00049-1

Deng, J., Dong, W., Socher, R., Li, L., Kai, L., and Li, F.-F. (2009). "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, 248–255.

Dodge, S., and Karam, L. (2017). "Can the early human visual system compete with Deep Neural Networks?," in *Proceedings of the IEEE International Conference on Computer Vision Workshop*, (Venice: IEEE), 2798–2804.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., et al. (1994). fMRI of human visual cortex. *Nature* 369, 525–525.

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* 41, 1149–1160. doi: 10.3758/brm.41.4.1149

Fawzi, A., Fawzi, O., and Frossard, P. (2018). Analysis of classifiers' robustness to adversarial perturbations. *Mach. Learn.* 107, 481–508. doi: 10.1007/s10994-017-5663-3

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. (2016). "Robustness of classifiers: from adversarial to random noise," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, (Barcelona: Curran Associates Inc), 1632–1640.

Fenske, M. J., Aminoff, E., Gronau, N., and Bar, M. (2006). "Chapter 1 Top-down facilitation of visual object recognition: object-based and context-based contributions," in *Progress in Brain Research*, eds S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J. M. Alonso, and P. U. Tse (Amsterdam: Elsevier), 3–21. doi: 10.1016/s0079-6123(06)55001-0

Flesch, T., Balaguer, J., Dekker, R., Nili, H., and Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proc. Natl. Acad. Sci.* 115, 10313–10322.

Geirhos, R., Temme, C. R. M., Rauber, J., Schuett, H. H., Bethge, M., and Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1808.08750 (accessed February 23, 2021).

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014a). "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, (Cambridge, MA: MIT Press), 2672–2680.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1412.6572 (accessed February 23, 2021).

Grill-Spector, K., and Malach, R. (2004). The human visual cortex. *Annu. Rev. Neurosci.* 27, 649–677.

Güçlü, U., and van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/jneurosci.5023-14.2015

Güçlü, U., and van Gerven, M. A. J. (2017). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *Neuroimage* 145, 329–336. doi: 10.1016/j.neuroimage.2015.12.036

Hong, H., Yamins, D. L., Majaj, N. J., and Dicarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19, 613–622. doi: 10.1038/nn.4247

Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* 8:15037.

Jozwik, K. M., Kriegeskorte, N., and Mur, M. (2016). Visual features as stepping stones toward semantics: explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia* 83, 201–226. doi: 10.1016/j.neuropsychologia.2015.10.023

Kang, X., Yund, E. W., Herron, T. J., and Woods, D. L. (2007). Improving the resolution of functional brain imaging: analyzing functional data in anatomical

space. *Magn. Reson. Imaging* 25, 1070–1078. doi: 10.1016/j.mri.2006.12.005

Kay, K., Jamison, K. W., Vizioli, L., Zhang, R., Margalit, E., and Ugurbil, K. (2019). A critical assessment of data quality and venous effects in sub-millimeter fMRI. *Neuroimage* 189, 847–869. doi: 10.1016/j.neuroimage.2019.02.006

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., and Wandell, B. A. (2013). GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* 7:247. doi: 10.3389/fnins.2013.00247

Kay, K. N., and Yeatman, J. D. (2017). Bottom-up and top-down computations in word-and face-selective cortex. *eLife* 6:e22341.

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., and Mcdermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron* 98, 630–644. doi: 10.1016/j.neuron.2018.03.044

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., and Kriegeskorte, N. (2017). Fixed versus mixed RSA: explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* 76, 184–197. doi: 10.1016/j.jmp.2016.10.007

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1312.6114 (accessed February 23, 2021).

Knill, D. C., and Pouget, A. (2004). The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 25, 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P., and Soatto, S. (2017). Analysis of universal adversarial perturbations. *arXiv* [Preprint]. Available online at: https://arxiv.org/pdf/1705.09554.pdf (accessed February 23, 2021).

Needell, D., and Vershynin, R. (2009). Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Computat. Math.* 9, 317–334. doi: 10.1007/s10208-008-9031-3

Nguyen, A., Yosinski, J., and Clune, J. (2015). "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Boston, MA: IEEE), 427–436.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computat. Biol.* 10:e1003553. doi: 10.1371/journal.pcbi.1003553

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178. doi: 10.1038/nn.3495

Rahnev, D. (2017). Top-down control of perceptual decision making by the prefrontal cortex. *Curr. Direct. Psychol. Sci.* 26, 464–469. doi: 10.1177/0963721417709807

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and Dicarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* 38, 7255–7269. doi: 10.1523/jneurosci.0388-18.2018

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annu. Rev. Vis. Sci.* 5, 399–426. doi: 10.1146/annurev-vision-091718-014951

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1312.6199 (accessed February 23, 2021).

Ullman, S., Assif, L., Fetaya, E., and Harari, D. (2016). Atoms of recognition in human and computer vision. *Proc. Natl. Acad. Sci. U.S.A.* 113, 2744–2749.

Vinje, W. E., and Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287, 1273–1276. doi: 10.1126/science.287.5456.1273

Wandell, B. A., Dumoulin, S. O., and Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron* 56, 366–383. doi: 10.1016/j.neuron.2007.10.012

Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., and Liu, Z. (2017). Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* 28, 4136–4160. doi: 10.1093/cercor/bhx268

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and Dicarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yamins, D. L. K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365. doi: 10.1038/nn.4244

Yuan, X., He, P., Zhu, Q., Bhat, R. R., and Li, X. (2017). Adversarial examples: attacks and defenses for deep learning. *arXiv* [Preprint]. Available online at: https://arxiv.org/abs/1712.07107 (accessed February 23, 2021).

Zanto, T. P., Rubens, M. T., Thangavel, A., and Gazzaley, A. (2011). Causal role of the prefrontal cortex in top-down modulation of visual processing and working memory. *Nat. Neurosci.* 14:656. doi: 10.1038/nn.2773

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., and Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *Neuroimage* 32, 1826–1836. doi: 10.1016/j.neuroimage.2006.04.199

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 01 September 2021
doi: 10.3389/fninf.2021.667375

Medical Image Interpolation Using Recurrent Type-2 Fuzzy Neural Network

Jafar Tavoosi[1], Chunwei Zhang[2]*, Ardashir Mohammadzadeh[3], Saleh Mobayen[4]* and Amir H. Mosavi[5,6]*

[1] Department of Electrical Engineering, Ilam University, Ilam, Iran, [2] Structural Vibration Control Group, Qingdao University of Technology, Qingdao, China, [3] Department of Electrical Engineering, University of Bonab, Bonab, Iran, [4] Future Technology Research Center, National Yunlin University of Science and Technology, Douliou, Taiwan, [5] Faculty of Civil Engineering, Technische Universität Dresden, Dresden, Germany, [6] Institute of Software Design and Development, Obuda University, Budapest, Hungary

Image interpolation is an essential process for image processing and computer graphics in wide applications to medical imaging. For image interpolation used in medical diagnosis, the two-dimensional (2D) to three-dimensional (3D) transformation can significantly reduce human error, leading to better decisions. This research proposes the type-2 fuzzy neural networks method which is a hybrid of the fuzzy logic and neural networks as well as recurrent type-2 fuzzy neural networks (RT2FNNs) for advancing a novel 2D to 3D strategy. The ability of the proposed methods in the approximation of the function for image interpolation is investigated. The results report that both proposed methods are reliable for medical diagnosis. However, the RT2FNN model outperforms the type-2 fuzzy neural networks model. The average squares error for the recurrent network and the typical network reported 0.016 and 0.025, respectively. On the other hand, the number of fuzzy rules for the recurrent network and the typical network reported 16 and 22, respectively.

Keywords: recurrent neural network, type-2 fuzzy system, image interpolation, 2D to 3D, brain MRI, artificial intelligence, machine learning

OPEN ACCESS

Edited by:
Tolga Cukur,
Bilkent University, Turkey

Reviewed by:
Mehmet Saadeddin Öztürk,
Karadeniz Technical University, Turkey
Ugurhan Kutbay,
Gazi University, Turkey

****Correspondence:***
Jafar Tavoosi
j.tavoosi@ilam.ac.ir
Chunwei Zhang
zhangchunwei@qut.edu.cn
Saleh Mobayen
mobayens@yuntech.edu.tw

Received: *12 February 2021*
Accepted: *11 August 2021*
Published: *01 September 2021*

Citation:
Tavoosi J, Zhang C, Mohammadzadeh A, Mobayen S and Mosavi AH (2021) Medical Image Interpolation Using Recurrent Type-2 Fuzzy Neural Network. Front. Neuroinform. 15:667375. doi: 10.3389/fninf.2021.667375

INTRODUCTION

In medical imaging, a cross-sectional sequence of high-resolution organs or tissues is obtained using CT, MRI, or other methods (Leng et al., 2013). However, the distance between neighboring slices is usually much larger than the pixel size, which is attributed to the ability of imaging devices or time/storage/dose constraints (Neuberta et al., 2012). The direct use of such data for three-dimensional (3D) image reconstruction often results in inaccurate images due to the heterogeneous dimensions of the images, the structure of discontinuous errors, sharp points, and other errors. To obtain volumetric (3D) data with isotropic dimensions and to reconstruct the 3D structure, it is essential to conduct several interpolations between the sections (Pan et al., 2012). In the other words, as the conventional imaging devices are two-dimensional (2D), to have a 3D accurate image for better diagnosis and treatment, a set of 2D images are often taken and combined (Ebied et al., 2018). However, one of the major problems is the presence of blind or undefined dots in one or more of the images. To address this issue a 2D interpolation operation is used (Hung et al., 2019). Recently, several methods for 2D interpolation have been proposed. Some of which are discussed as follows. In (Leng et al., 2013), while expressing the problem of various categories of image

interpolation methods, multiple resolution methods have been used to internalize medical images. In this method, first, a few images are taken with different resolutions and then by internalizing them, a small image is extracted but with minimal information deletion. Classical mathematical techniques use a set of basic functions to estimate the values between cuts. The nearest neighbor, B-Spline linear, and cube functions are standard types of these techniques. Such introspection approaches are commonly used in modern medical imaging (Neuberta et al., 2012). To improve the accuracy, different families of spline functions have been accepted and used as introspection cores (Pan et al., 2012). The interpolation operation itself leads to an increase in image size, but in order to store images, they must first be reduced in size so that they do not take up much space (Ebied et al., 2018). In research, interpolation methods are divided into two categories: scene-oriented and goal-oriented. Scene-based techniques are effective and easy to implement but can produce significant artifacts that relate pixels that occupy the same matrix location in continuous images to different anatomical structures. But in contrast to target-based introspection techniques, the information in the image slices is used to facilitate more accurate introspection (Hung et al., 2019). The second category is much more common and has received more attention because, for example, in an image of the brain, a mass can be considered as a target and a 3D view of the target can be obtained more accurately. In (Triwijoyo and Adil, 2021), the Bicubic interpolation algorithm is used to resize images and then by analyzing the three parameters of mean square error, mean square root of error, and maximum signal-to-sound ratio and analysis and They analyze the superiority of their work over the Bilinear method and nearest neighbor algorithms have shown. They also concluded that Bilinear and Nearest-neighbors increase the level of computational complexity (Murad et al., 2021). Presents a method based on efficient interpolated compressed sensing to increase the speed and accuracy of MRI devices. In (Iglesias et al., 2021), the deep convolutional neural network (NN) is used to internalize medical images. In the mentioned paper, the effect of variable contrasts and different orientations is considered and acceptable results are obtained. The disadvantages of the mentioned paper are time-consuming and an average square error of more than 0.01.

Today, computational intelligence has permeated most sciences (Tavoosi et al., 2011c, 2017a; Pour Asad et al., 2016, 2017). Artificial NNs have been extensively employed in the medical sciences, especially in predictive discussion (Maihami et al., 2016; Kazemi et al., 2017; Ayat, 2018; Armand et al., 2019; Tabatabaei et al., 2019). However, not much has been done in the field of interpolation of medical images using computational intelligence methods (neural network, fuzzy logic, etc.). In the following, some of the work done in this field will be reviewed. In (Chao and Kim, 2019), the fuzzy neural network of the radial base function has been used to internalize medical images. Accordingly, two suspended images are normally used to be inserted as the input of the fuzzy NN. The final output data is obtained using a learned NN. In this article, 6 entries and 3 membership functions are considered for each. Therefore, the number of fuzzy rules is $3^6 = 243$. Meanwhile, the number of neurons in the third and fourth layers is 4,374 and 729, respectively, which complicates the network, and also the execution time of the program will be very long. Naturally, such a structure will not be able to run online. A comparative plan for the development of core-based introspection methods that simultaneously improves image resolution and maintains accurate local edges is presented in Chen and Wang (2010).

Medical imaging researchers have been inspired by the advancement of deep learning methods and computational resources to combine deep learning in medical image analysis. Some recent studies have shown that accurate algorithms are successfully used to segment medical imaging and diagnose and classify diseases. A deep learning method is presented in Havaei et al. (2017), in which the network uses a circular layer instead of a fully connected layer to accelerate the segmentation process. A cascading structure is used, which compares the output of the first network with the successful network input. The network provided in Pereira et al. (2016) uses small cores to classify pixels in the image. Using small cores, without worrying about over-training (Sharifian et al., 2011), reduces the number of network parameters and helps to create deeper networks (Tavoosi et al., 2011a,b; Tavoosi et al., 2012). Increasing and normalizing its intensity has been done in the preprocessing phase to facilitate the training process. Using fuzzy theory with a genetically based learning algorithm, first the distance corresponding to the new edge is estimated based on local slope information, and then this estimated distance is used in different introspection methods instead of the main Euclidean distance. In addition, a genetic algorithm-based learning method is provided to automatically obtain important parameters of the fuzzy system. In short, for a particular pixel being processed, we replace the six input pixels. In turn, we can obtain basic data to produce an integrated interpolation image. Subsequently, several upgrade cycles were applied to train NN. Finally, we can get the desired output through the updated neural network (Tavoosi et al., 2016a,b; Tavoosi and Azami, 2019). In (Deepika et al., 2021), a fuzzy neural network has been used to compress medical images and read them quickly. The deep neural network has been used to classify and quickly read medical images (Puttagunta and Ravi, 2021). In the mentioned paper, a convolutional neural network (2D) has been used, which has a high speed, but unfortunately, it does not have good accuracy and has some errors. Type-2 fuzzy logic (T2FL) is rarely used in image interpolation, and this method is still in its infancy. For example, recently in Mohammed and Hussain (2021) a combination of Mamdani T2FL with a convolutional neural network has been used to identify images of animals. Although the mentioned article has some drawbacks such as not examining different angles, not examining blurry images, not examining the background of the same color as the animal, etc.,. However, because it is at the beginning of the path, it is generally appropriate and acceptable. In this paper, we propose a new method based on a RT2FNN. This structure consists of five layers (Tavoosi et al., 2017b). In the following, the problem will be explained first. Then we talk about the RT2FNN. In continue, the evaluation of the proposed method is presented by simulation and at the end, the conclusion is expressed.

STATEMENT OF THE PROBLEM

According to the sequence of images (called the cut) $\{I_i\}_{i=0}^{N}$ with the same size $(w + 1)\ x\ (h + 1)$; where w and h are two positive integers; **Figure 1**). The problem of image embedding to determine the sequence of preserved properties between intersections $J_{i,m}, i = 0, ..., N_1, m = 1, 2, ..., M_i$ so that $\bigcap_{m=1}^{M_i} \{J_{m,i}\}$. Continuous transfer from I_i to I_{i+1}. Here M_i is the number of cuts inserted between I_i and I_{i+1}. For simplicity, assume that M_i for i = 0.1,..., N–1. The inputs are discrete matrices. To facilitate calculations, the matrix $I = \left[I_{ij} \right]_{(w+1)\times(h+1)}$ with size $(w + 1) \times (h + 1)$ can be used as $I(u)$. With $u = [u.v]^T \in \Omega = [0, 1]^2$. Here, $I\left(\frac{i}{w}, \frac{i}{h}\right) = I_{ij}$, $i = 0, 1, ..., w, j = 0, 1, ..., h$ and other values of the function are calculated by two-line (bipolar) interpolation. We define the set of points $U = \{u_{ij} = \left[\frac{i}{w}, \frac{i}{h}\right]^T : i = 0, 1, ..., w, j = 0, 1, ..., h\}$ as the set of pixel points. The embedding techniques are applied through the scene or object approches. For scene-based interpolation, the internal image quality values are extracted from the image quality values given in the same situations. Simple scene-based methods are easy to calculate but may results in remarkable artifacts (waste or noise). As you know, the images obtained from the detection are very linear, blurry, and obscure.

Since the information of shape and structure are not used, then the recording-based technique can be considered as an object-based technique. Note, for example, that the two images $I_o(u)$ and $I_1(u)$, the recording-based theme always consists of two steps; first, using a recording method to draw an image with another image, second, to create a deformation, a traditional attachment is employed.

The basic idea of image capture is to find an $X(u)$ conversion that adapts one image to another. The similarity of the two images is considered to be the sum of the differences in square intensity (SSD), cross-correlation, and cross-sectional information. Converting $X(u)$ can be hard or non-hard. Hard conversion is easier, and there are fewer parameters such as transmission, rotation, and scalability. However, the non-hard transfer is more flexible. $X(u)$ transmission is usually represented by a fragmentary-linear function. To smooth out the transfer, some settings minimize the changes in the minor derivatives of the transfer. Therefore, the functional energy for the model is written as follows.

$$\mathcal{E}\,(\mathrm{x}) = \mathcal{S}\left(I_0\left(x\left(u\right)\right), I_1\left(u\right)\right) + \mathcal{R}\left(x\right) \tag{1}$$

The first semester measures the similarity between $I_0(X(u))$ and $I_1(u)$, and the second semester represents the order of $X(u)$. In the similarity statement, I_0 is deformed by X, while I_1 is constant. Here, the relation (1) that transforms only one image is called the one-way model.

Note that the roles of I_0 and I_1 are symmetrical in the one-way model. Suppose that the $x_0(u)$ map for the I_0 deformation is made in such a way that $I_o(x_0(u)) \approx I_1(u)$. Furthermore, another map $x_1(u)$ is constructed to meet the conditions $I_1(x_1(u)) \approx I_0(u)$. This process is called back recording. It should be noted that the hole The loop in I_1 cannot appear when I_0 is converted to I_1. When I_1 is converted to I_0, the holes become smaller. Maps x_0 and x_1 are displayed by the B_spline function and they are displayed using grids. The network map $x(u)(u = [u, t]^T \in [0, 1]^2)$ from a vertical curve family $x(\frac{i}{20}, v)_{i=0}^{20}$; and a set of horizontal curves $x(u, \frac{i}{20})_{j=0}^{20}$ is formed. If the record is backward, and the reverse trend is facing to be forward, the equation $x_0 = x_1^{-1}$ must be or estimate $x_0^{-1} = x_1$. However, we can see that forward recording is not the opposite of posterior recording, and therefore the role of I_0 and I_1 in the single-directional recording, the model is different. $J_{mid}^{(01)}(u)$ is in the middle of the middle image, which is inserted based on the front recording, and the image $J_{mid}^{(01)}(u)$ is in the middle of the middle image obtained by the rear recording. The derived images through the backward/forward recording are a general way to prevent saturation. However, the artificial effect (parasite) may still exist, as the images created by the two recording approaches may be quite different from each other. In the suggested method, the reshaping of both I_0 and I_1 are considered.

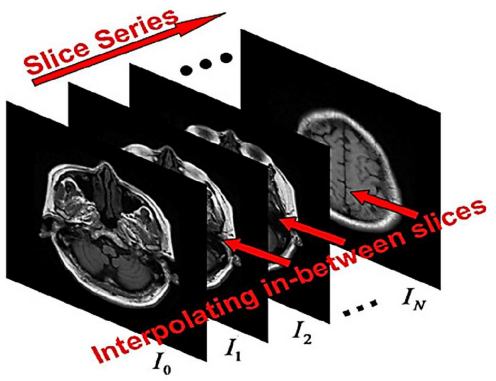


FIGURE 1 | The sequence of cross-sections to create a three-dimensional image.

Methods and Design of Algorithms

There are three steps you can take to begin the process of preparation for mediation. The suggested recording and introspection approaches are described in this section.

According to the two images I_0 and I_1 with size $(w + 1) \times (h + 1)$, we use them as continuous functions $I_0(u)$ and $I_1(u)$ and

$$u = [u, t]^T \in \Omega[0, 1]^2 \tag{2}$$

Introducing bipolar introspection. In the registration model, we are going to find two maps:

$$x_k = [x_k, y_k]^T : \Omega \to \Omega, k = 0, 1 \tag{3}$$

In which case the following is established:

(i) x_k are C^2 maps :

(ii)$x_k\,(0, v) = 0, x_k\,(1, v) = 1, y_k\,(u, 0) = 0$ and $y_k\ (u, 1) = 1$;

(iii) for a given $\in (0 < \varepsilon < 1)$,

$$det\ [x_{ku}, x_{kv}] \geq \in \text{ on } \Omega.$$

where in,

$$\begin{aligned}\mathcal{E}(x_0, x_1) = \int_\Omega \frac{[I_0(x_0(u)) - I_1 x_1(u))]^2}{1.0 + c[I_0(x_0(u))^2 + I_1(x_1(u))^2]} du \\ +\lambda_1 \sum_{k=0}^{1} \int_\Omega \left[\| x_{ku}(u) \|^2 + \| x_{kv}(u)) \|^2\right] du \\ +\lambda_2 \sum_{k=0}^{1} \int_\Omega \| x_{ku}(u) \times x_{kv}(\mathrm{u}) \|^2 du\end{aligned} \tag{4}$$

Equation (4) must be minimized. Here $x_{ku}(u)$ and $x_{kv}(u)$ express the first partial derivatives of $x_k(u)$ relative to the variables u and v. The condition (iii) is said to be a regular condition that guarantees x_k as an injection map. Set $(V(\Omega) = X : \Omega \to \Omega$; justifying conditions 1 to 3) Then $\forall x \in V(\Omega)$ is an internal map one by one.

In this paper, the maps $x_k(u)$ $(k = 0,1)$ are expressed as two-variable cube elliptical functions with vector B-spline of the size defined in $\prime\Omega$. The first term of the energy function (4) refers to the similarity relationship used to minimize the error between the two deformed images. Regarding the inconvenience of registration appraoh, some constraints are needed to make $x_k(u)(k = 0, 1)$ as much as possible. The second and third relations are applied to smooth the transformations $x_k(u)(k = 0, 1)$We set $\mathcal{R}_1 \sum_{k=0}^{1} \int_\Omega [\| x_{ku}(u) \|^2 + \| x_{kv}(u)) \|^2] du$ two relationships Call the first time and go to $\mathcal{R}_2(x_0, x_1) \sum_{k=0}^{1} \int_\Omega (\|x_{ku} \times x_{kv} \|^2) du$ Let's say that the parameters λ_1 and λ_2 are two specific coefficients of regulatory expressions. The following is an interpretation and analysis of the model:

The Relationship of Similarity

The two maps $x_0(u)$ and $x_1(u)$ are designed to reshape I_0 and I_1 in such a way that the reshaped images of $I_k(x_k(u))\ (k = 0, 1)$ are similar. Compared to one-dimensional recording, the use of two maps in the same period overcomes the saturation problem I_0 and I_1. Also, images $I_0(x_u(u))$ and $I_1(x_1(u))$ match using only one map, because the free parameters has doubled.

To measure the similarity between the two images I_0 and I_1, a simple metric and a cheap computation, the SSD is $\int_\Omega (I_0(u) - I_1(u))^2 du$. However, the endurance of non-compliance for the area of difference between the intensity of the squares of the low-intensity area is greater than that of the more intense area. In practical applications, low-intensity features are as important as high-intensity features. In practical applications, low-intensity features are as important as high-intensity features. Therefore, a modified criterion $\int_\Omega g(u)(I_0(u) - I_1 u))^2 du$ (SSD) is applied to our model. The term $g(u) = 1/[1.0 + c(I_0(u)^2 + I_1(u)^2]$ is denoted by the denominator of at least 1 and c as a constant positive number.

The Relationship of the First Order Settings

For the desired mapping $x(u)$, $\int_0^1 \| x_u(u, v_0) \| du$ is the length of the arc curve $C(u) := x(u, v_0)$. Also $\int_0^1 \| x_v(u_0, v) \| dv$ and the length of the arc curve is $C(v) := x(u_v, v)$. Therefore, the setting of the first-order expression $\int_\Omega \| x_u(u)^2 \| + \| x_v(\mathrm{u}) \|^2 du$ intends to map the conversion of x according to the variables u and v. The phrase "regulator" denotes a convex function based on x, so we can get at least the predicate if and only if x is the same mapping.

Phrase Regional Settings

In parametric form $x : \Omega \longrightarrow \Omega$, the surface element is written by $\| x_u \times x_v \| du$. For Ω=[0, 1]^2, the area $x\varepsilon\ V(\Omega)$ is equal to: $|\Omega| = \int_\Omega \| x_u \times x_v \| du$=1. Using the inequality Kushi_Shuartz:

$1 = \int_\Omega \| x_u \times x_v \| du \le \left(\int_\Omega \| x_u \times x_v \|^2 du^{\frac{1}{2}}\right).||^{\frac{1}{2}}$, The equation is established if and only if $\| x_u \times x_v \| \equiv 1$. Thus, the relational constraints of the $\int_\Omega \| x_u \times x_v \|^2 du$ constraints the constraints so that the area element remains constant. Since $\| x_u \times x_v \|^2 = \| x_u \|^2 \| x_v \|^2 - < x_u, x_v >^2$, the relationship of the regional settings can be $\int_\Omega \left(\| x_u \|^2 \| x_v \|^2 - < x_u, x_v >^2\right) du$ also wrote.

Select the Parameters λ_1 and λ_2 in the First-Order Settings

The choice of parameters λ_1 and λ_2 depends on the deformation between the two specific images. The first-order setting term $\mathcal{R}_1$ monitors the flexibility of $x_k(k = 0, 1)$ conversions. Therefore, we have to consider λ_1 as large if the image has high strength, and vice versa, if we consider λ_1 to be small, there are many differences between I_0 and I_1. Relationship settings $||x_{ku} \times x_{kv}|| = \left|\det[x_{ku}, x_{kv}]\right|$ for $k = 0, 1$ Limits area elements (triple conditions). Therefore, with a small λ_2, the setting process may have to be stopped because the three conditions are not satisfactory. However, these images cannot match well with a large λ_2, because the elastic deformation is stopped by the $\mathcal{R}_2$ regulator. The registration approach is not very sensitive to regulatory parameters. In our experiments, the parameters λ_1 and λ_2 are experimentally

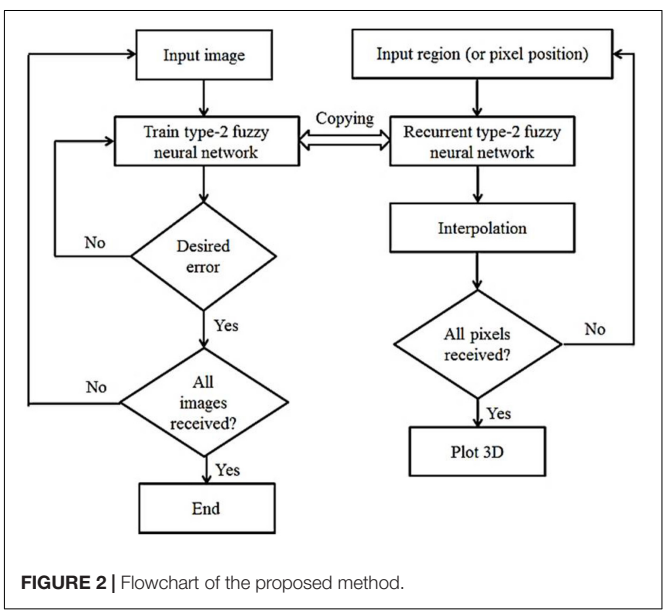


FIGURE 2 | Flowchart of the proposed method.

selected based on the given images. **Figure 2** shows the general flowchart of the work. In this figure, the left part is related to the learning phase of the recurrent type-2 neural network and the right part is related to the interpolation phase using it.

The procedure according to **Figure 2** is that first the existing images are entered into the system one by one, processed and the RT2FNN is trained. The condition for completing the training is to achieve the minimum desired error. Then, in the second phase, the blind spots or pixels that are not available are trained, generated and the interpolation operation is completed by the type 2 recursive neural network, and finally, if all the pixels are identified, a 3D image printing command is issued.

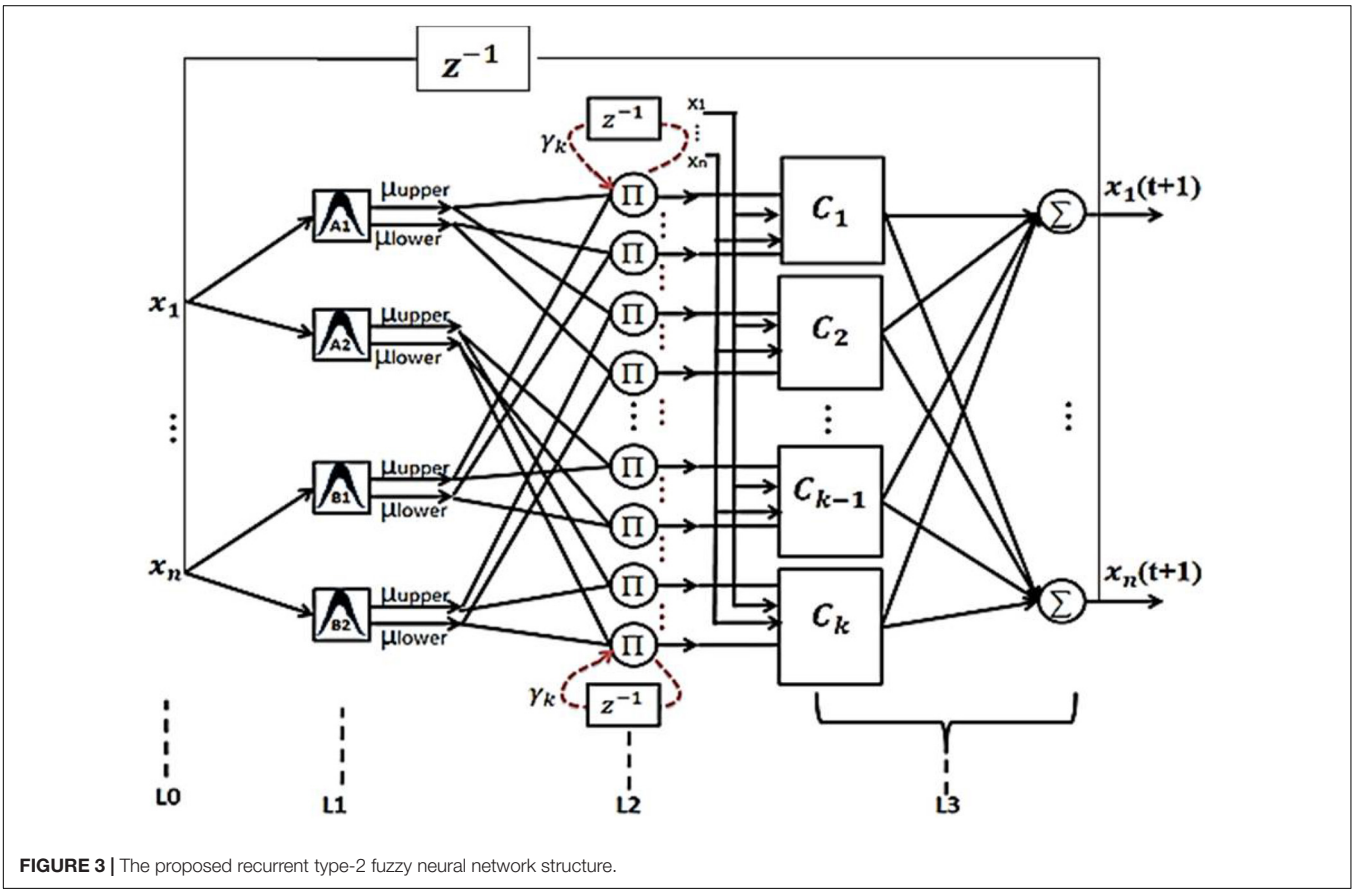


FIGURE 3 | The proposed recurrent type-2 fuzzy neural network structure.

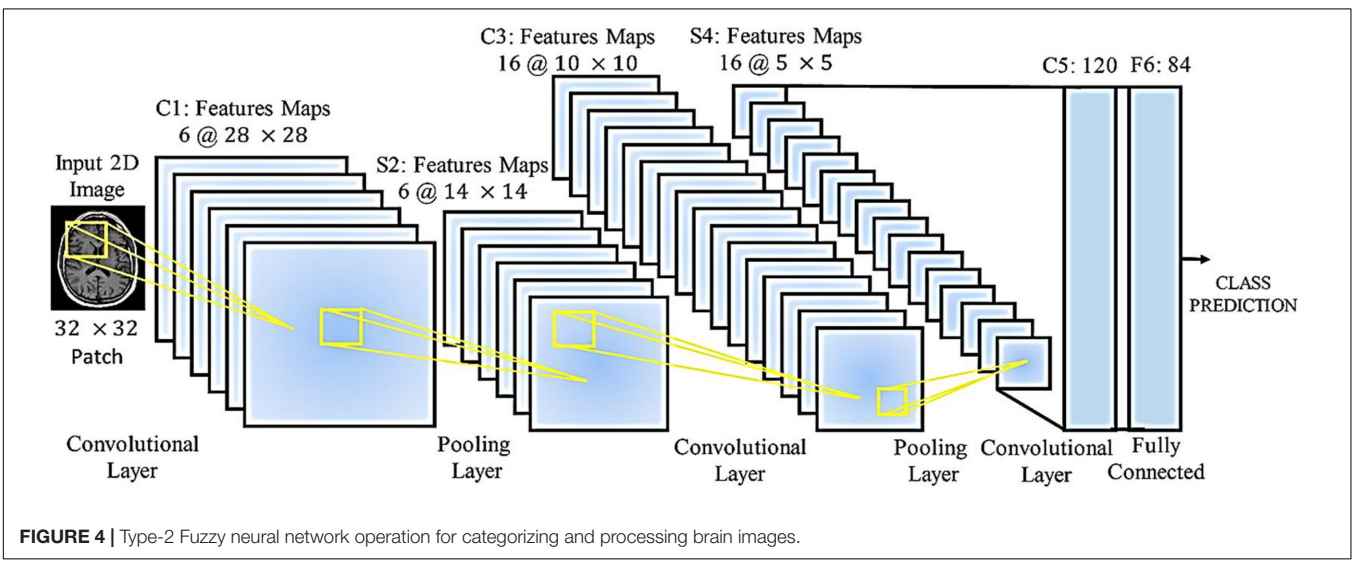


FIGURE 4 | Type-2 Fuzzy neural network operation for categorizing and processing brain images.

RECURRENT TYPE-2 FUZZY NEURAL NETWORK MODEL

The structure of the proposed RT2FNN is shown in **Figure 3**.

Details of how the proposed network works and how to train it are described in Tavoosi et al. (2017b). Based on a specific pixel of the recorded image, six pixels (3–3) are selected from the two conditional images. Each pixel is represented by 3 fuzzy subsets. Then, based on the condition formulation layer, a general fuzzy rule is written by the combination of 6 fuzzy subsets. We determine the rule. By random assignment, (6×729) 4,374 some conditions define all the relevant rules. The number of neurons in the formulation layer is 4374 and number of rules is 729. Finally, the output pixel is calculated. Sequentially, we can obtain all the output data needed to embed an image by processing the total recorded image.

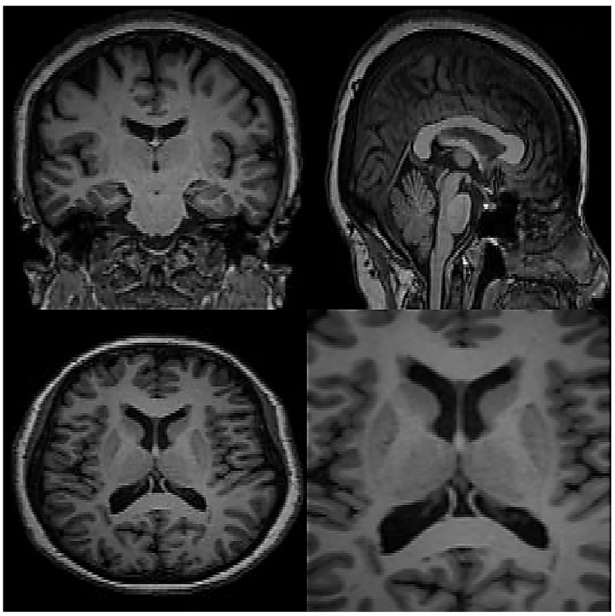


FIGURE 5 | The results of recurrent type-2 fuzzy neural network for two-dimensional interpolation.

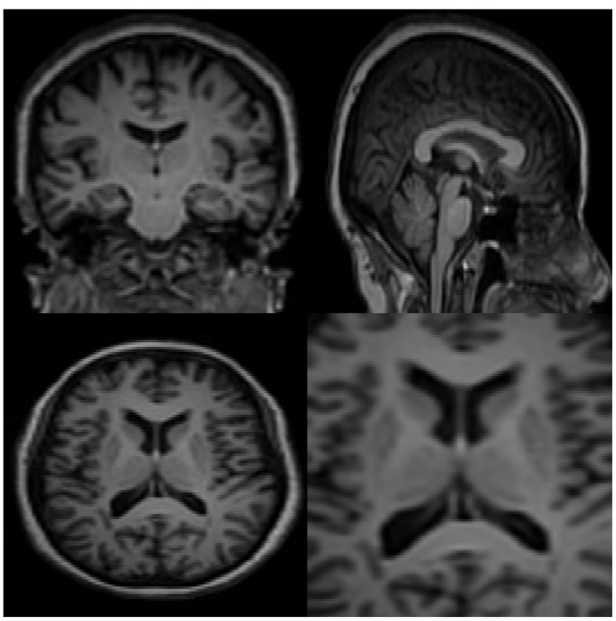


FIGURE 6 | The results of a typical type-2 fuzzy neural network for two-dimensional interpolation.

INTERPOLATION USING RECURRENT TYPE-2 FUZZY NEURAL NETWORK

A wide range of medical imaging techniques are used to predict and diagnose clinical problems, but in most cases, the images obtained are similar, with deep learning where the network structure allows this, gives us solutions. Once specialized knowledge is available in this field, then the manipulated features

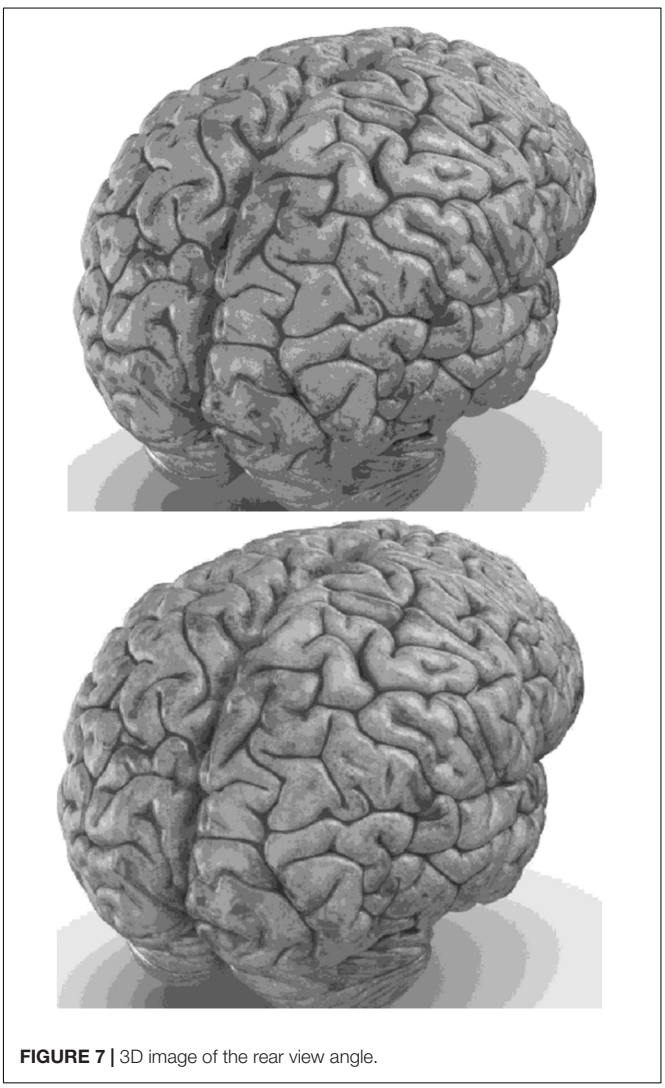


FIGURE 7 | 3D image of the rear view angle.

operate, and in general it can be said that this creates difficult and complex assumptions, and these assumptions may be for some. Do not use medical imaging. So despite the hand-made features, it's hard to tell the difference between healthy and unhealthy images in some cases. A classifier such as a support device (SVM) does not provide a final and comprehensive solution. Features derived from methods such as the criterion for converting immutable properties (SIFT) are independent of the task or task assigned. Classifiers such as vector support have been applied to this model, and no mechanism has been improved to lose local features, which in the process of extracting features and classifying those that are separated from each other, does not exist. On the other hand, a RT2FNN learns these properties through basic data. These features are guided data and are learned to end the learning mechanism. The ability of the fuzzy neural network to regenerate is that the error signal obtained in lost functions is extracted and reused to improve properties (fuzzy-type recursive fuzzy-type fuzzy network filters in layers primary are taught). Therefore, RT2FNNs become a better model. Another advantage is that in the early layers of a RT2FNN, the edges surround the spots and in the local structure, while the nerve cells in the upper layers focus more on the part. Different people have human organs, some of which can completely consider the human organs in the final layers. **Figure 4** shows the processing of brain images using a neural network.

Figure 4 shows the medical images for the classification of medical images by accepting a 32 × 32 N class fragment from the original 2D image. The network has loops, maximum volume, and fully connected layers. Each annular layer produces a linear design of different sizes, and the volume of the layers reduces the size of the linear designs to be transferred to the lower layers. Fully connected layers produce the prediction of the intended class at the output. Several parameters require a network, which depends on the number of layers, the number of nerve cells in each layer, and the relationship between these nerve cells.

The training phase of the network ensures that the maximum possible efficiency is learned in which the best performance is possible to solve the desired problem.

SIMULATION

In this section, the data for brain reconstruction images are first extracted from the Allen Brain Atlas Database. The size of the images can be from 256 × 256 pixels to more than 4,000 × 4,000 pixels. Our algorithm can work with any size. The larger the size,

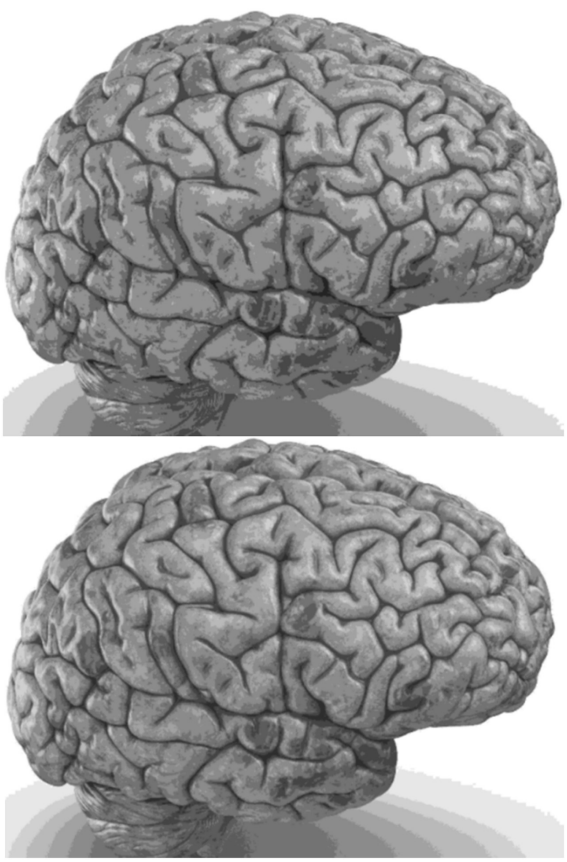


FIGURE 8 | 3D image from the right viewing angle.

FIGURE 9 | 3D image from the left viewing angle.

the longer the processing time, but the higher the accuracy. In this article, we have used 100 images of 768 × 578 (442 kB) to create 3D images. The neural network has 5 intermediate layers, each layer having 100 neurons. Then, these 2D images were used to teach the RT2FNN and normal one. **Figure 3** shows the results of RT2FNNs for 2D interpolation. **Figure 5** shows the results of a typical type-2 fuzzy neural network for 2D interpolation. Note that a normal (typical) network does not have feedback, i.e., it does not use past moment data. In fact, one of the purposes of this article is to show the importance and impact of the presence or absence of feedback in the structure of the neural network.

The specifications of the computer users are as follows: Windows 10 Home 64; 11th Gen Intel® Core™ i7 processor; Intel® Iris® Xe Graphics; 8 GB memory; 256 GB Intel® SSD Storage; 16 GB Intel®. The simulation was performed in MATLAB software version 2019a.

Carefully in **Figures 5**, **6**, it can be seen that the recurrent network has performed better, especially in detail. In the following, the ability of the recurrent network and the normal network to reconstruct the 3D image of the brain is examined. The 3D image of the brain is depicted from four angles. **Figure 7**, which is shown from a rearview angle, the above figure is the results from the recurrent network and the bottom figure is the results from the normal network. Also shown in **Figures 8–10** are 3D images of the brain from the right, left, and front viewing angles, respectively. In all images, the above figure is for the recurrent network and the bottom figure is for the typical network.

TABLE 1 | Specifications of networks used.

	Number of fuzzy rules	Average squares error	Training time
Recurrent network	16	0.016	570 s
Typical network	22	0.025	510 s
Method of Neuberta et al. (2012)	–	0.15	4 s

FIGURE 10 | 3D image from the front viewing angle.

DISCUSSION

The main purpose of this paper is to show the importance of feedback and the use of past moment data in the structure of type-2 fuzzy neural networks. Of course, for the first time, these networks have been used for interpolation in medical images, and this is another innovation of this paper. Naturally, the more accurate the 3D image, the easier it is for doctors and relevant specialists to work with the least error. Looking at **Figures 5–10**, it can be seen that the accuracy of recurrent networks is objectively higher than normal networks, and the reason for this is that recurrent networks use the information of neighboring points in the images. In **Figures 8–10**, as can be seen in blind spots or deep spots (indent), the difference between a recurrent and a normal network becomes more pronounced. This is especially true in **Figure 10**, as this image has many dents and blind spots. The reason for the superiority of the RT2FNN over conventional is the simultaneous use of output and input data to the network, while in the conventional network only input information is used. In other words, the return network has a dynamic structure, but the normal network operates statically. The network specifications used are shown in **Table 1**.

As can be seen in **Table 1**, the number of fuzzy rules of the recurrent network is less than the typical network. The average error squares in the return network are far less than the typical type. But the training time of recurrent networks is longer than the training time of the typical network, which is due to the existence of feedback and its calculations. For further comparison, we also used methods without the use of computational intelligence (fuzzy logic, neural network, etc.). For example, you can see the results of the method presented in Neuberta et al. (2012) in the table. The average squares error

is much higher than the intelligence-based methods, but the processing time is surprisingly short. In general, time can be sacrificed for accuracy, because a high-quality 3D image is more needed by the medical community than the time it takes to create and produce this image.

CONCLUSION

In this study, the problem of interpolation in medical images using RT2FNNs was addressed. Image interpolation is used for two main purposes. Firstly, to increase the quality of images adding the number of pixels is studied. Secondly, 3D images are produced. The recurrent type-2 fuzzy neural networks model outperforms the type-2 fuzzy neural networks model. The average squares error for the recurrent network and the typical network reported 0.016 and 0.025, respectively. On the other hand, the number of fuzzy rules for the recurrent network and the typical network reported 16 and 22, respectively. The recurrent type-2 fuzzy neural network has internal feedback, and as it uses output information it can therefore provide more accurate interpolation with less error. It is worth mentioning that the training time of the images for the recurrent type-2 fuzzy neural networks model is longer. However, in the medical sciences, model accuracy is more important. It is expected that the methodology represented in this research would be extended further in 3D printers, tumor surgeries, and so on. For future studies using type-3 fuzzy logic and color images will be considered.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://openfmri.org/dataset/.

ETHICS STATEMENT

The Ethics Committee of Ilam University approved the study.

AUTHOR CONTRIBUTIONS

JT: writing-original draft preparation, software, validation, visualization, investigation, and supervision. CZ and AM: writing-original draft editing, visualization, investigation, and supervision. AM: writing-original draft editing, software, validation, visualization, and investigation. SM: writing-original draft editing, validation, software, and validation. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by Ilam University.

ACKNOWLEDGMENTS

Open Access Funding by the Publication Fund of the TU Dresden.

REFERENCES

Armand, R., Rigi, G., and Bahrami, T. (2019). Fuzzy hybrid least-squares regression approach to estimating the amount of extra cellular recombinant protein A from *Escherichia coli* BL21. *J. Ilam Univ. Med. Sci.* 27, 1–13. doi: 10.29252/sjimu.27.3.1

Ayat, S. (2018). Increasing the speed and precision of prediction of the results of angiography by using combination of adaptive neuro-fuzzy inference system and particle swarm optimization algorithm based on data from Kowsar Hospital of Shiraz. *J. Ilam Univ. Med. Sci.* 26, 142–154. doi: 10.29252/sjimu.26.4.142

Chao, Z., and Kim, H. J. (2019). Slice interpolation of medical images using enhanced fuzzy radial basis function neural networks. *Comput. Biol. Med.* 10, 66–78. doi: 10.1016/j.compbiomed.2019.05.013

Chen, H. C., and Wang, W. J. (2010). Locally edge-adapted distance for image interpolation based on genetic fuzzy system. *Expert Syst. Appl.* 37, 288–297. doi: 10.1016/j.eswa.2009.05.069

Deepika, J., Rajan, C., and Senthil, T. (2021). Security and privacy of cloud- and IoT-based medical image diagnosis using fuzzy convolutional neural network. *Comput. Intell. Neurosci.* 2021, 1–17. doi: 10.1155/2021/6615411

Ebied, M., Elmisery, F. A., and Zekry, A. (2018). "Utilization of decimation interpolation strategy for medical image communication and storage," in *Proceedings of the 2018 8th International Conference on Computer Science and Information Technology (CSIT)*, Amman, 22–25.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., et al. (2017). Brain tumor segmentation with deep neural networks. *Med. Image Anal.* 35, 18–31.

Hung, K. W., Wang, K., and Jiang, J. (2019). Image interpolation using convolutional neural networks with deep recursive residual learning. *J. Multimed. Tools Appl.* 78, 22813–22831. doi: 10.1007/s11042-019-7633-1

Iglesias, J. E., Billot, B., Balbastre, Y., Tabari, A., Conklin, J., González, R. G., et al. (2021). Joint super-resolution and synthesis of 1 mm isotropic MP-RAGE volumes from clinical MRI exams with scans of different orientation, resolution and contrast. *Neuroimage* 237:118206. doi: 10.1016/j.neuroimage.2021.118206

Kazemi, M., Mehdizadeh, H., and Shiri, A. (2017). Heart disease forecast using neural network data mining technique. *J. Ilam Univ. Med. Sci.* 25, 20–32. doi: 10.29252/sjimu.25.1.20

Leng, J., Xu, G., and Zhan, Y. (2013). Medical image interpolation based on multi-resolution registration. *Comput. Math. Appl.* 66, 1–18. doi: 10.1016/j.camwa.2013.04.026

Maihami, V., Khormehr, A., and Rahimi, E. (2016). Designing an expert system for prediction of heart attack using fuzzy systems. *Sci. J. Kurdistan Univ. Med. Sci.* 21, 118–131.

Mohammed, H. R., and Hussain, Z. M. (2021). Hybrid mamdani fuzzy rules and convolutional neural networks for analysis and identification of animal images. *Computation* 9:35. doi: 10.3390/computation9030035

Murad, M., Jalil, A., Bilal, M., Ikram, S., Ali, A., Khan, B., et al. (2021). Radial undersampling-based interpolation scheme for multislice CSMRI reconstruction techniques. *Biomed Res. Int.* 2021:6638588.

Neuberta, A., Salvado, O., Acosta, O., Bourgeat, P., and Fripp, J. (2012). Constrained reverse diffusion for thick slice interpolation of 3D volumetric MRI images. *Comput. Med. Imaging Graph.* 36, 130–138. doi: 10.1016/j.compmedimag.2011.08.004

Pan, M. S., Yang, X. L., and Tang, J. T. (2012). Research on interpolation methods in medical image processing. *J. Med. Syst.* 36, 777–807. doi: 10.1007/s10916-010-9544-6

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in mri images. *IEEE Trans. Med. Imaging* 35, 1240–1251. doi: 10.1109/tmi.2016.2538465

Pour Asad, Y., Shamsi, A., Ivani, H., and Tavoosi, J. (2016). Adaptive intelligent inverse control of nonlinear systems with regard to sensor noise and parameter uncertainty (magnetic ball levitation system case study). *Int. J. Smart Sens. Intell. Syst.* 9, 148–169. doi: 10.21307/ijssis-2017-864

Pour Asad, Y., Shamsi, A., and Tavoosi, J. (2017). Backstepping-based recurrent type-2 fuzzy sliding mode control for MIMO systems (MEMS triaxial gyroscope case study). *Int. J. Uncertain. Fuzziness Knowl. Based Syst.* 25, 213–233.

Puttagunta, M., and Ravi, S. (2021). Medical image analysis based on deep learning approach. *Multimed. Tools Appl.* 80, 24365–24398.

Sharifian, M. B. B., Mirlo, A., Tavoosi, J., and Sabahi, M. (2011). "Self-Adaptive RBF Neural Network PID Controller in Linear Elevator," in *Proceedings of the International Conference on Electrical Machines and Systems*, Beijing.

Tabatabaei, S. M. R., Saadatjoo, F., and Mirzaei, M. (2019). The prediction model for cardiovascular disease using Yazd's health study data (YaHS). *J. Shahid Sadoughi Univ. Med. Sci.* 27, 1346–1360.

Tavoosi, J., Alaei, M., and Jahani, B. (2011a). "Temperature Control of Water Bath by Using Neuro-Fuzzy Controller," in *Proceedings of the 5th Symposium on Advance in Science and Technology*, Mashhad.

Tavoosi, J., Alaei, M., Jahani, B., and Daneshwar, M. A. (2011b). A novel intelligent control system design for water bath temperature control. *Aust. J. Basic Appl. Sci.* 5, 1879–1885.

Tavoosi, J., and Azami, R. (2019). A new method for controlling the speed of a surface permanent magnet synchronous motor using fuzzy comparative controller with hybrid learning. *J. Comput. Intell. Electr. Eng.* 10, 57–68.

Tavoosi, J., Badamchizadeh, M. A., and Ghaemi, S. (2011c). Adaptive inverse control of nonlinear dynamical system using type-2 fuzzy neural networks. *J. Control* 5, 52–60.

Tavoosi, J., Shamsi Jokandan, A., and Daneshwar, M. A. (2012). A new method for position control of a 2-DOF robot arm using neuro-fuzzy controller. *Indian J. Sci. Technol.* 5, 2253–2257.

Tavoosi, J., Suratgar, A. A., and Menhaj, M. B. (2016a). Nonlinear system identification based on a self-organizing type-2 fuzzy RBFN. *Eng. Appl. Artif. Intell.* 54, 26–38. doi: 10.1016/j.engappai.2016.04.006

Tavoosi, J., Suratgar, A. A., and Menhaj, M. B. (2016b). Stable ANFIS2 for nonlinear system identification. *Neurocomputing* 182, 235–246. doi: 10.1016/j.neucom.2015.12.030

Tavoosi, J., Suratgar, A. A., and Menhaj, M. B. (2017a). Stability analysis of recurrent type-2 TSK fuzzy systems with nonlinear consequent part. *Neural Comput. Appl.* 28, 47–56. doi: 10.1007/s00521-015-2036-3

Tavoosi, J., Suratgar, A. A., and Menhaj, M. B. (2017b). Stability analysis of a class of MIMO recurrent type-2 fuzzy systems. *Int. J. Fuzzy Syst.*19, 895–908. doi: 10.1007/s40815-016-0188-7

Triwijoyo, B. K., and Adil, A. (2021). Analysis of medical image resizing using bicubic interpolation algorithm. *J. Ilmu Komput.* 14, 20–29.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 20 October 2021
doi: 10.3389/fninf.2021.676491

Classification of Obsessive-Compulsive Disorder Using Distance Correlation on Resting-State Functional MRI Images

Qian Luo[1,2,3†], *Weixiang Liu*[1,2,3†], *Lili Jin*[4,5], *Chunqi Chang*[1,2,3,6*] *and Ziwen Peng*[5,7*]

[1] School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China, [2] Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, Shenzhen University, Shenzhen, China, [3] National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Shenzhen University, Shenzhen, China, [4] Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou, China, [5] Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China, [6] Peng Cheng Laboratory, Shenzhen, China, [7] Department of Child Psychiatry, Shenzhen Kangning Hospital, Shenzhen University School of Medicine, Shenzhen, China

Both the Pearson correlation and partial correlation methods have been widely used in the resting-state functional MRI (rs-fMRI) studies. However, they can only measure linear relationship, although partial correlation excludes some indirect effects. Recent distance correlation can discover both the linear and non-linear dependencies. Our goal was to use the multivariate pattern analysis to compare the ability of such three correlation methods to distinguish between the patients with obsessive-compulsive disorder (OCD) and healthy control subjects (HCSs), so as to find optimal correlation method. The main process includes four steps. First, the regions of interest are defined by automated anatomical labeling (AAL). Second, functional connectivity (FC) matrices are constructed by the three correlation methods. Third, the best discriminative features are selected by support vector machine recursive feature elimination (SVM-RFE) with a stratified N-fold cross-validation strategy. Finally, these discriminative features are used to train a classifier. We had a total of 128 subjects out of which 61 subjects had OCD and 67 subjects were normal. All the three correlation methods with SVM have achieved good results, among which distance correlation is the best [accuracy = 93.01%, specificity = 89.71%, sensitivity = 95.08%, and area under the receiver-operating characteristic curve (AUC) = 0.94], followed by Pearson correlation and partial correlation is the last. The most discriminative regions of the brain for distance correlation are right dorsolateral superior frontal gyrus, orbital part of left superior frontal gyrus, orbital part of right middle frontal gyrus, right anterior cingulate and paracingulate gyri, left the supplementary motor area, and right precuneus, which are the promising biomarkers of OCD.

Keywords: obsessive-compulsive disorder, functional connectivity, distance correlation, classification, rs-fMRI

OPEN ACCESS

Edited by:
Sharlene D. Newman, University of Alabama, United States

Reviewed by:
Feng Liu, Tianjin Medical University General Hospital, China
Shihui Ying, Shanghai University, China

***Correspondence:**
Chunqi Chang
cqchang@szu.edu.cn
Ziwen Peng
pengzw@m.scnu.edu.cn

[†]These authors have contributed equally to this work

Received: *05 March 2021*
Accepted: *13 September 2021*
Published: *20 October 2021*

Citation:
Luo Q, Liu W, Jin L, Chang C and Peng Z (2021) Classification of Obsessive-Compulsive Disorder Using Distance Correlation on Resting-State Functional MRI Images. Front. Neuroinform. 15:676491. doi: 10.3389/fninf.2021.676491

INTRODUCTION

Obsessive-compulsive disorder (OCD) is a mental disorder that causes repeated and unwanted thoughts and/or obsessive feelings and compulsive actions and it can limit the ability of the patient to take part in relationships, the workplace, and in society (Piacentini et al., 2003; Abramowitz et al., 2009). Its prevalence is about 1–3% lifetime (Ruscio et al., 2010; Rapinesi et al., 2019). In

clinical practice, no diagnostic biomarkers are available for OCD and its diagnosis is always based on some symptom-oriented criteria according to the International Classification of Diseases (ICD; Stein et al., 2016) and the Diagnostic and Statistical Manual of Mental Disorders (DSM; Battle, 2013). However, these criteria may have several problems over the conditions of an individual. For example, the patients with OCD often co-occur with depression and anxiety or another psychiatric comorbidity, which can contribute to misdiagnosis.

With the development of medical imaging, researchers can explore the pathogenesis of OCD. Currently, the pathogenesis of OCD has been confirmed to be caused by the cortico-striato-thalamo-cortical (CSTC) circuit dysfunction, but emerging evidence indicates that broader brain regions, such as the left supplementary motor area (SMA) and right precuneus, are involved in this disorder (Saxena et al., 1998; Rehn et al., 2018; Thorsen et al., 2018; Hazari et al., 2019). These changes in the brain are due to the diversity of tasks in the investigation of OCD. Therefore, task-based functional MRI (task-fMRI) has been studied for detecting the functional changes in the brain in patients with OCD and their relatives (Menzies et al., 2008a). However, task-fMRI studies can only focus on some specific regions of the brain and may have missed important information existing in regions of the brain not related to the task. Without specific design in task-fMRI, resting-state functional MRI (rs-fMRI) provides an effective and noninvasive approach to assess the neural activation and functional connectivity (FC) of the human brain without any hypothesis. It can also provide a reliable measure of baseline brain activity and may complement and extend findings from task-based studies (Biswal et al., 1995; Hou et al., 2014; de Vries et al., 2019; Yang et al., 2019).

Recently, the multivariate pattern analysis based on a machine learning (ML) algorithm has been introduced for neuroimaging analysis of a variety of diseases such as autism, depression, and schizophrenia (Sajda, 2006; Anderson et al., 2011; Zeng et al., 2012; Liu et al., 2014, 2015a, 2017; Mueller et al., 2015; Rathore et al., 2017; Lamothe et al., 2018; Zhou et al., 2018; Bu et al., 2019; Rapinesi et al., 2019). It has the advantage of being able to inference individual level over the univariate analysis used at the group level (Orrù et al., 2012; Goodman et al., 2014). In comparison to other traditional methods of analysis, its ability to use inter-regional correlations, such as the Pearson correlation, to detect subtle and spatially distributed effects (Menzies et al., 2008b; Bruin et al., 2020; Zhan et al., 2021). Therefore, it seems particularly well-suited for the neuroimaging analyses in OCD, as abnormalities are typically distributed across the brain (Klöppel et al., 2008; Arbabshirani et al., 2017).

In this study, we employed the multivariate pattern analysis *via* the three correlation methods to distinguish the patients with OCD from a healthy control subject (HCS). A general flowchart of rs-fMRI based on the FC matrix for diagnosis is shown in **Figure 1**. In this framework, there are four main steps: (1) defining the region of interests (ROIs) from the rs-fMRI images or by using the anatomically and functionally defined reference atlases of the brain, (2) extracting rs-fMRI time series based on the ROIs and calculating the FC matrices, (3) using feature selection method to get the optimal features from the FC matrices, and (4) training a classifier.

Currently, this study mainly focused on the second and third parts. In the second part, with rs-fMRI time series data, the FC matrix can be extracted for characterizing the network structure of the brain. One way is to calculate the Pearson correlation between rs-fMRI time series over the ROIs predefined as automated anatomical labeling (AAL) with 116 structural regions (Tzourio-Mazoyer et al., 2002). For example (Shenas et al., 2013; Gruner et al., 2014; Sen et al., 2016; Takagi et al., 2017), the authors use Pearson correlation as the network features. In addition, the partial correlation was also used for measuring the FC (Varoquaux et al., 2010; Smith et al., 2011; Dadi et al., 2019). However, the Pearson correlation and partial correlation only discover the linear dependency, although the partial correlation excludes the indirect influence of the correlation structure. To overcome this limitation, the distance correlation was proposed to measure both linear and non-linear associations between the two ROIs (Szekely et al., 2007; Yoo et al., 2019).

In the third part, because of high-dimensional features from the FC matrix, we need a feature selection algorithm to reduce the dimensionality. In literature, the recursive feature elimination (RFE) algorithm is a very excellent feature selection technique that has been widely used in many fields (Guyon et al., 2002; Ding et al., 2015; Liu et al., 2015b; Lin et al., 2017; Wang et al., 2019), but it needs a specific classifier. Currently, studies on the diagnosis of OCD disease are usually limited to a small data set, so the researchers tend to use traditional ML methods to complete the task. Among them, support vector machine (SVM) provides excellent performance (Shenas et al., 2013; Gruner et al., 2014; Sen et al., 2016; Takagi et al., 2017; Wang et al., 2019). Therefore, we applied the SVM-RFE algorithm to filter the features.

In the fourth part, these selected features were entered into the seven classifiers. According to the final classification performance, we can explore the optimal FC method and classifier and investigate the regions of the brain, which may be potential biomarkers. Finally, our aims were 2-fold: one is to investigate which correlation method achieves the best discrimination between OCD and HCS and the other is to explore some potential biomarkers according to the above results.

MATERIALS AND METHODS

Participants

This study was approved by the Ethics Committee of Shenzhen Kangning Hospital and the written informed consent was obtained from each participant. A total of 128 subjects were enrolled from Shenzhen Kangning Hospital and Guangzhou Brain Hospital, including 67 HCS and 61 patients with OCD, aged from 13 to 63 years old. The demographic information and clinical characteristics information are shown in **Table 1**. The independent sample *t*-test was carried out on the age ($p = 0.45$). For the Yale-Brown Obsessive Compulsive Scale (Y-BOCS) total score, the Y-BOCS obsessions score, and the Y-BCOS compulsions score ($p < 0.001$), we used the independent sample Kruskal–Wallis test.

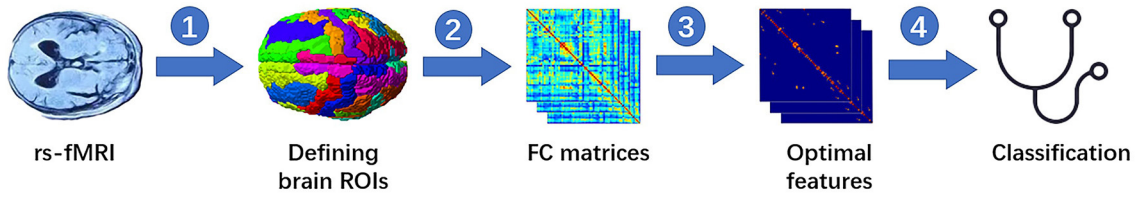


FIGURE 1 | Classification of flow chart with four main steps: (1) defining the ROIs from the rs-fMRI images or by using anatomical templates that have been defined, (2) extracting rs-fMRI time series based on the ROIs and calculating the FC matrices, (3) using feature selection method to define the optimal features from the FC matrices, and (4) training a classifier. ROIs, region of interests; rs-fMRI, resting-state functional MRI; FC, functional connectivity.

TABLE 1 | Demographic and clinical characteristics of the participants.

Variable	OCD	HCS	*p*-value
Demographic measure			
Number	61	67	-
Sex	M43, F18	M 48, F 19	-
Age average	27.7 ± 8.4	28.9 ± 8.5	$p = 0.45$
Clinical measures			
YBOCS total score	26.8 ± 6.1	2.3 ± 3.4	$p < 0.001$
YBOCS obsessions score	14.6 ± 3.8	1.2 ± 1.7	$p < 0.001$
YBCOS compulsions score	12.2 ± 5.0	1.2 ± 2.1	$p < 0.001$

OCD, obsessive-compulsive disorder; HCS, healthy control subject; Y-BOCS, Yale-Brown Obsessive Compulsive Scale.

Imaging Data Acquisition

A 3.0-Tesla MR system (Philips Medical Systems, Best, the Netherlands) equipped with an eight-channel phased-array head coil was used for the data acquisition. Functional data were collected by using gradient echo-planar imaging (EPI) sequences [time repetition (TR) = 2,000 ms, echo time (TE) = 60 ms, flip angle = 90°, 33 slices, field of view (FOV) = 240 mm^2 × 240 mm^2, matrix = 64 × 64, slice thickness = 4 mm, and voxel size = 3.75 mm^3 × 3.75 mm^3 × 4 mm^3]. For each participant, the fMRI scanning lasted for 480 s and 240 volumes were obtained. For spatial normalization and localization, a high-resolution T1-weighted anatomical image was also acquired by using a magnetization prepared gradient echo sequence (TR = 8 ms, TE = 3.7 ms, flip angle = 7°, FOV = 240 mm^2 × 240 mm^2, matrix = 256 × 256, slice thickness = 1 mm, and voxel size = 0.94 mm^3 × 0.94 mm^3 × 1 mm^3). During the scanning, the participants were instructed to relax with their eyes closed and stay awake without moving.

Data Preprocessing

The data were preprocessed by using the Statistical Parametric Mapping toolbox (SPM12, https://www.fil.ion.ucl.ac.uk/spm) and the Data Processing Assistant for Resting-State fMRI (DPARSF version 4.4, http://rfmri.org/dpabi; Shenas et al., 2013, 2014). Image preprocessing consisted of: (1) removing first the 10-time points; (2) slicing timing correction; (3) realigning the time series of the images for each subject; (4) T1-weighted individual structural images by coregistered to the mean functional image; (5) the transformed structural images by segmented into gray matter, white matter, and cerebrospinal fluid; (6) based on these segmented images, using diffeomorphic anatomical registration through exponentiated lie algebra (DARTEL) (Ashburner, 2007) tool to estimate the normalization parameters from individual native space to the Montreal Neurological Institute (MNI) space (Xue et al., 2020); (7) the functional imaging data normalized to the MNI space by using these normalization parameters and resampling at 3 mm^3 × 3 mm^3 × 3 mm^3; (8) nuisance covariate regression (head motion parameters, white matter signal, and cerebrospinal fluid signal); (9) spatial smoothing with a 4-mm full-width half-maximum isotropic Gaussian kernel; (10) band-pass filtering (0.01–0.08 Hz); and (11) micro-head-motion correction according to framewise displacement (FD) by replacing the rs-fMRI volume with FD > 0.5 mm (nearest neighbor interpolation).

Definition of ROIs and Calculation of the FC Matrix

In this study, we employed an AAL atlas to define the ROIs. For the calculation of the FC matrix, the Pearson correlation, partial correlation, and distance correlation methods will be used in this study. For each subject, the mean of time series over all voxels in each region was extracted. The FC matrices were calculated between these average time courses with the three correlation methods implemented with Nilearn software (http://nilearn.github.io/). Considering that the matrix was symmetric, we only needed to take the lower triangle of the matrix. Finally, we flattened the lower trig matrix to get a feature vector with a length of (116 × 116 − 116)/2 = 6,670. In our following experiment,

each feature was normalized by Fisher's z-transformation (Fisher, 1915; Vergun et al., 2013; Kassraian-Fard et al., 2016).

Feature Selection and Classification

In this study, to reduce the dimension of the data and find the most discriminative subset of feature, we applied SVM-RFE with a stratified N-fold cross-validation strategy for feature selection (SVM-RFE-NCV). It is a sequential backward selection algorithm based on the maximum margin principle of SVM under the N-fold cross-validation. The process contains five steps: (1) training the model with the samples, (2) sorting the scores of each feature, (3) removing the features with the minimum scores, (4) training the model again with the remaining features and repeating the process, and (5) selecting the required features (Ding et al., 2015; Wang et al., 2019). With the selected features, seven classifiers are compared: SVM with linear kernel, multilayer perceptron (MLP), extreme gradient boosting (XGBoost), gradient boosting decision tree (GBDT), graph convolution network (GCN), and sparse L1 and non-sparse L2 regularization for the logistic regression classifiers (LR-L1 and LR-L2; Friedman et al., 2001; Chen and Guestrin, 2016; Kipf, 2017). The SVM-RFE-NCV process was embedded in a classification framework with 10-fold cross-validation (10-CV).

Performance Evaluation

The performance of the proposed classifiers is assessed by using the four performance measures: specificity, sensitivity, accuracy, and area under the receiver-operating characteristic curve (AUC). To test whether these classification scores are significant, we performed a permutation test: we first randomly reassigned the subject labels and then performed the 10-CV classification. This procedure was repeated by 1,000 times. The p-value was then calculated by dividing the number of times that showed a higher value than the derived from the non-permuted model by the total number of permutations (Plitt et al., 2015).

RESULTS

In this study, some qualitative and quantitative comparison results are provided. At first, we qualitatively compared the three methods *via* the scatter plot and correlation visualization. Then, the classification results of the OCD and HCS are evaluated according to the pipelines composed of the three correlation measures, the SVM-RFE-NCV, and the seven classifiers, to select correlation measure and classifier to obtain the best discrimination between the OCD and HCS. In addition, we will use the SVM-RFE-NCV to find the regions of the brain corresponding to the most discriminative features. The SVM-RFE-NCV and classification algorithms are implemented by using Scikit-learn (Pedregosa et al., 2012).

Scatter Plot and Correlation Visualization

To explore the differences among the Pearson, partial, and distance correlations, we calculated the average functional matrices of the patients and HCS, respectively. Since the distance correlation coefficient ranges from 0 to 1 while the other two range from−1 to 1, we used the unsigned versions of the Pearson and partial correlation coefficients (for example, taking the absolute value of them). The results are shown in **Figure 2** and we can see that both the distance and Pearson correlations give similar structures of the functional matrix, while the structure of partial correlation is greatly different from them. To further reflect the similarities and differences among the Pearson, distance and partial correlations, we draw their scatter plots as shown in **Figure 3**. The values of three coefficients mainly lie in the different intervals: (-0.1, 0.9) for the Pearson correlation coefficients, (-0.1, 0.1) for the partial correlation coefficients, and (0.1, 0.8) for the distance correlation coefficients.

Choice of FC Method

To find the optimal correlation method, we proceeded in two steps. First, for each correlation method, we used the SVM-RFE-NCV to find the best feature subset that gave the prediction. Second, we compared the performance of each correlation method on the best feature subset.

Best Feature Subset

For the SVM-RFE-NCV, the number of optimal features (NOFs) varies with N. **Table 2** summarizes the changes of accuracy and NOF under the different N conditions. For the Pearson correlation, partial correlation, and distance correlation, the performance is the highest when N is equal to 5, 8, and 5, respectively. Therefore, the best feature subset of the Pearson correlation and distance correlation was obtained by the SVM-RFE-5CV algorithm. The best feature subset of partial correlation was achieved by the SVM-RFE-8CV algorithm.

Best Correlation Method

Three correlation methods produced a good performance in the classification. Their ROC curves are shown in **Figure 4**, from which we can see that they exhibit good performance, with AUC values range from 0.87 to 0.94 ($p < 0.01$). The other classification results of the three correlation methods for the patients with OCD and HCS are summarized in **Table 3**. The distance correlation and Pearson correlation are slightly lower than partial correlation in sensitivity, but distance correlation was the best in accuracy, specificity, and AUC followed by the Pearson correlation and partial correlation. Therefore, in the classification of the OCD and HCS, distance correlation comprehensive performance is the best. Its accuracy, sensitivity, and specificity are 93.01, 89.71, and 95.08% ($p < 0.01$), respectively. The second is Pearson correlation (accuracy = 89.74%, sensitivity = 89.71%, and specificity = 86.62%, $p < 0.01$). For partial correlation, accuracy is 84.87%, sensitivity is 96.21%, and specificity is 75.90% ($p < 0.01$).

Results of Different Classifiers

Through the above analysis, the best discriminative features can be obtained by using the distance correlation and the SVM-RFE-5CV. As stated earlier, the seven classifiers including SVM with linear kernel, MLP, XGBoost, GBDT, GCN, LR-L1, and LR-L2 classifiers were applied to identify these features separately. For SVM, the penalty parameter was set to 1. For LR-L1 and LR-L2, the penalty parameter was set to 0.01. For XGBoost, the learning

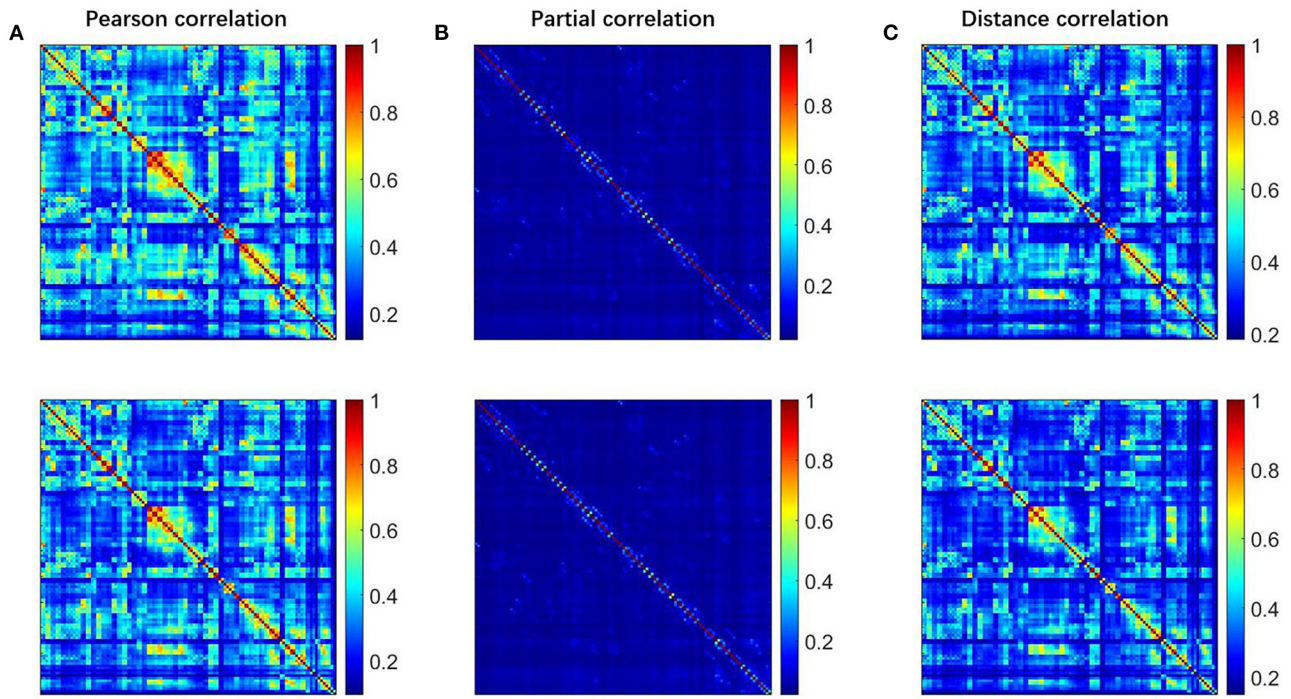


FIGURE 2 | The FC matrices for HCS and the patients with OCD. The first row is for HCS and the second row is for the patients with OCD. **(A)** Pearson correlation, **(B)** partial correlation, and **(C)** distance correlation. FC, functional connectivity; OCD, obsessive-compulsive disorder; HCS, healthy control subject.

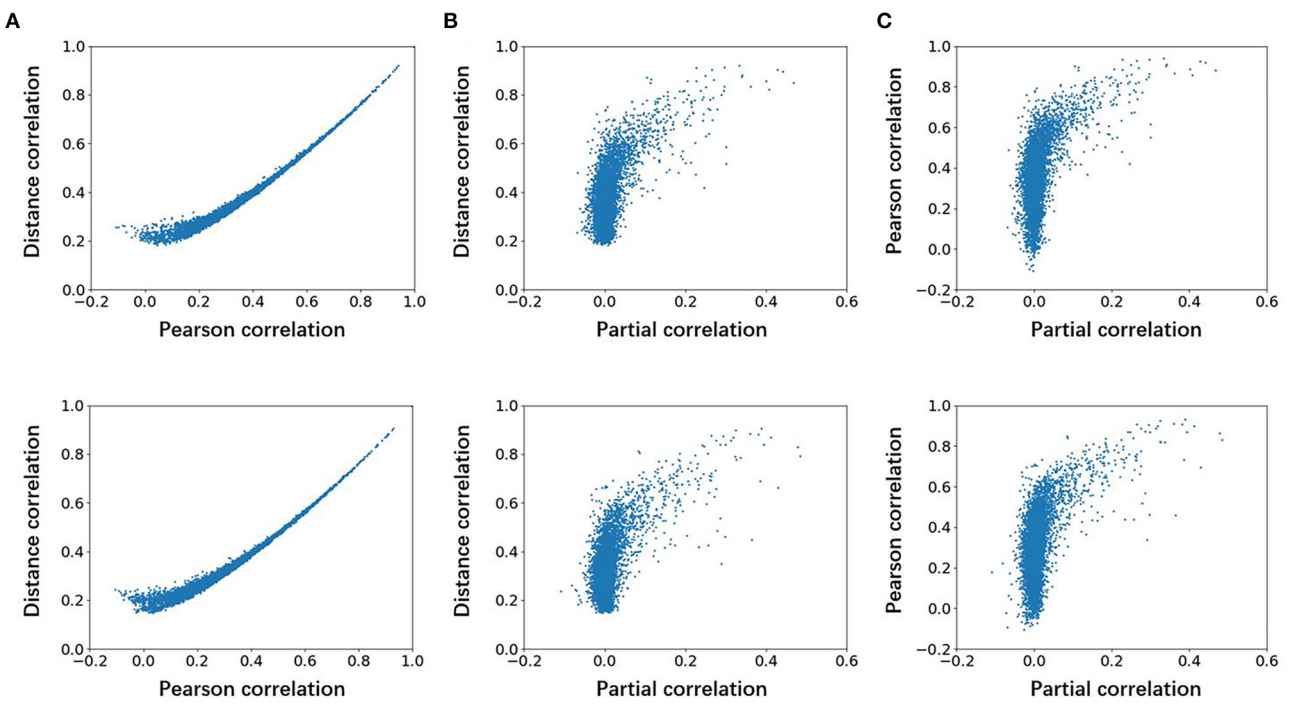


FIGURE 3 | Scatter plots of association among the Pearson, partial, and distance correlations for average functional connectivity across the participants. The first row is for HCS and the second row is for the patients with OCD. **(A)** Distance correlation vs. Pearson correlation, **(B)** distance correlation vs. partial correlation, and **(C)** pearson correlation vs. partial correlation. OCD, obsessive-compulsive disorder; HCS, healthy control subject.

TABLE 2 | Results of classification by the different number of the features.

Method	*N* = 3	*N* = 4	*N* = 5	*N* = 6	*N* = 7	*N* = 8	*N* = 9	*N* = 10
Pearson Correlation								
Accuracy (%)	85.77	89.68	89.74	88.91	87.31	87.37	85.76	88.89
NOF	108	83	84	233	111	119	101	106
Partial Correlation								
Accuracy (%)	80.13	77.82	80.06	79.29	80.71	84.87	81.60	80.83
NOF	896	829	1,489	1,150	971	2,488	1,467	1,531
Distance Correlation								
Accuracy (%)	85.06	89.03	93.01	89.87	86.60	88.91	89.80	87.43
NOF	60	91	81	96	107	245	112	412

NOF, number of features.

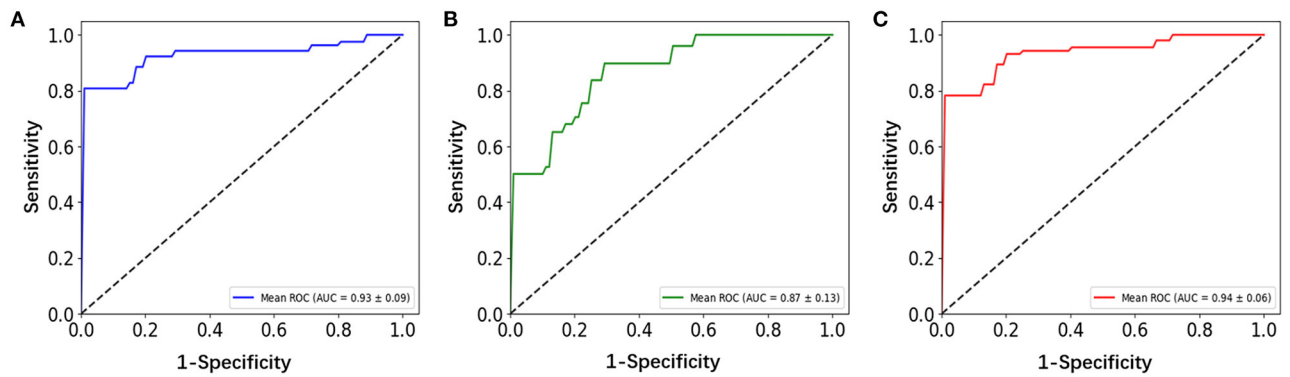


FIGURE 4 | ROC curves assessing Pearson correlation, partial correlation, and distance correlation performance by using SVM. **(A)** Pearson correlation, **(B)** partial correlation, and **(C)** distance correlation. ROC, receiver-operating characteristic; SVM, support vector machine.

TABLE 3 | The classification results of the three correlation methods.

Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	NOF
Pearson correlation	89.74 ± 7.28	89.71 ± 9.22	86.62 ±10.50	0.93 ± 0.09	84 ±46
Partial correlation	84.87 ± 7.09	96.21 ± 5.83	75.90 ± 17.33	0.87 ± 0.13	2488± 1393
Distance correlation	93.01 ± 5.40	89.71 ± 9.22	95.08 ± 7.70	0. 94 ± 0.06	81 ± 31

AUC, area under the receiver-operating characteristic curve; NOF, number of features.

rate was set to 0.01, the number of gradients boosted trees (n_estimators) to 200, maximum depth of the tree (max_depth) to 5, subsample ratio of the training instance (subsample) to 0.85, the minimum sum of instance weight needed in a child to 2, subsample ratio of the columns when constructing each tree to 0.7, and the other parameters to the default values. For GBDT, the learning rate was set to 0.01, n_estimators to 600, max_depth to 3, subsample ratio to 0.7, the minimum number of samples required to be at a leaf node to 10, the minimum weighted fraction of the total of weights required to be a leaf node to 0.1, and other parameters to the default values. For the GCN and MLP, dropout was set to 0.1, weight decay to 1×10^{-3}, learning rates to 0.02 and 0.05, number of epochs to 1,000, number of layers to 2, and the numbers of neurons per layer to 128 and 256. The results of classification by 10-CV are given in **Table 4**. The optimal classification result is achieved *via* SVM for accuracy, sensitivity, specificity, and AUC with values as high as 93.01, 89.71, 95.08%, 0.94 ($p < 0.01$), respectively.

Potential Biomarkers From Connectivity Patterns

To find the regions of the brain that strongly contributed to the discrimination between the patients with OCD and HCS, we selected the top 10 most discriminative features according to the SVM-RFE-NCV method. Specific regions of the brain were then located based on these features. The spatial maps of the regions of the brain (Xia et al., 2013)

TABLE 4 | Results of classification for the data of OCD.

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
SVM	93.01 ± 5.40	89.71 ± 9.22	95.08 ± 7.70	0.94 ± 0.06
LR-L1	89.81 ± 6.11	88.46 ± 10.23	91.47 ± 9.25	0.92 ± 0.07
LR-L2	90.58 ± 5.89	89.71 ± 9.22	91.29 ± 7.48	0.94 ± 0.06
GCN	91.41 ± 5.37	89.71 ± 9.22	92.72 ± 7.64	0.95 ± 0.06
MLP	90.64 ± 6.83	89.71 ± 9.22	91.29 ± 7.48	0.94 ± 0.06
XGBoost	85.77 ± 8.85	87.78 ± 11.19	84.84 ± 17.02	0.90 ± 0.12
GBDT	88.97 ± 7.23	86.71 ± 12.72	93.12 ± 9.49	0.94 ± 0.05

OCD, obsessive-compulsive disorder; SVM, support vector machine; LR-L1, sparse L1 for logistic regression; LR-L2, non-sparse L2 regularization for logistic regression; GCN, graph convolution network; MLP, multilayer perceptron; XGBoost, extreme gradient boosting; GBDT, gradient boosting decision tree; AUC, area under the receiver-operating characteristic curve.

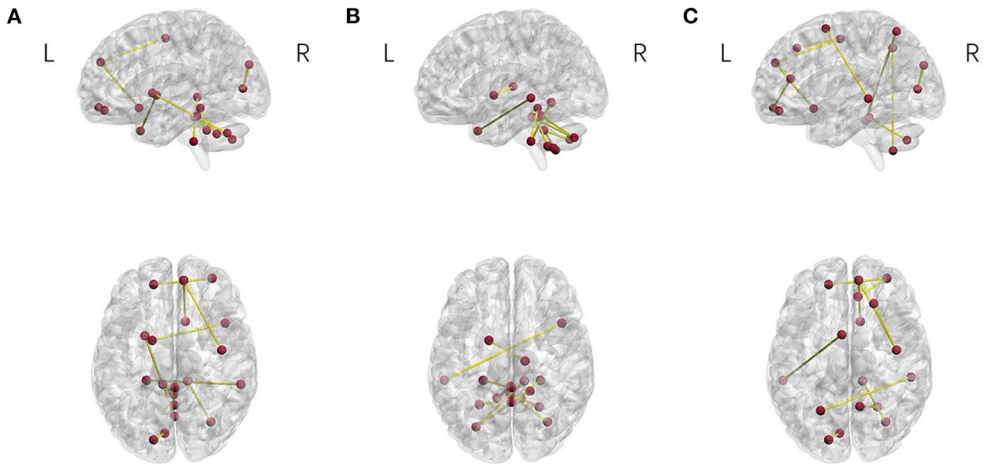


FIGURE 5 | Brain connectivity patterns for the three correlation methods. **(A)** Pearson correlation, **(B)** partial correlation, and **(C)** distance correlation.

are shown in **Figure 5** and the detailed information is listed in **Table 5**.

For Pearson correlation, the most discriminative regions included the right precentral gyrus, orbital part of left superior frontal, orbital part of right middle frontal gyrus, right olfactory cortex, the medial part of right superior frontal gyrus, left calcarine fissure and surrounding cortex, left superior occipital gyrus, left putamen, left globus pallidus, right middle temporal gyrus, right middle temporal pole, right crus II of cerebellar hemisphere, left lobule III of cerebellar hemisphere, right lobule III of cerebellar hemisphere, left lobule X of cerebellar hemisphere, lobule III of the vermis, lobule VIII of the vermis, lobule IX of the vermis, and lobule X of the vermis.

For partial correlation, the most discriminative regions for OCD were composed of the left globus pallidus, right thalamus, left middle temporal gyrus, right middle temporal pole, left crus II of cerebellar hemisphere, right crus II of cerebellar hemisphere, right lobule III of cerebellar hemisphere, right lobule IV of cerebellar hemisphere, right lobule V of cerebellar hemisphere, left lobule VIII of cerebellar hemisphere, right lobule VIII of cerebellar hemisphere, left lobule VIII of cerebellar hemisphere, right lobule IX of cerebellar hemisphere, left lobule X of cerebellar hemisphere, and right lobule X of the cerebellar hemisphere.

For distance correlation, the discriminative regions for OCD primarily consisted of the right percental gyrus, right dorsolateral superior frontal gyrus, orbital part of left superior frontal gyrus, orbital part of right middle frontal gyrus, left SMA, right olfactory cortex, the medial part of right superior frontal gyrus, right anterior cingulate and paracingulate gyri, left calcarine fissure and surrounding cortex, left superior occipital gyrus, left superior parietal gyrus, right precuneus, right superior temporal pole, right inferior temporal gyrus, left crus II of cerebellar hemisphere, left lobule III of cerebellar hemisphere, and right lobule VIII of the cerebellar hemisphere.

DISCUSSION

The goal of this study is to investigate the potential diagnostic value of the different correlation methods in patients with OCD. We systematically compare the FC matrix-based prediction methods. Our results show that the distance correlation method is optimally followed by the Pearson correlation and partial

TABLE 5 | The most discriminative brain regions.

Number	Pearson correlation	Partial correlation	Distance correlation
1	Right precentral gyrus	Left globus pallidus	Right percental gyrus
2	Orbital part of left superior frontal	Right thalamus	Right dorsolateral superior frontal gyrus
3	Orbital part of right middle frontal gyrus	Left middle temporal gyrus	Orbital part of left superior frontal gyrus
4	Right olfactory cortex	Right middle temporal pole	Orbital part of right middle frontal gyrus
5	Medial part of right superior frontal gyrus	Left crus II of cerebellar hemisphere	Left supplementary motor area
6	Left calcarine fissure and surrounding cortex	Right crus II of cerebellar hemisphere	Right olfactory cortex
7	Left superior occipitalgyrus	Right Lobule III of cerebellar hemisphere	Medial part of right superior frontal gyrus
8	Left putamen	Right lobule IV, V of cerebellar hemisphere	Right anterior cingulate and paracingulate gyri
9	Left globus pallidus	Left lobule VIII of cerebellar hemisphere	Left calcarine fissure and surrounding cortex
10	Right middle temporal gyrus	Right lobule VIII of cerebellar hemisphere	Left superior occipitalgyrus
11	Right middle temporal pole	Right lobule VIII of cerebellar hemisphere	Left superior parietal gyrus
12	Right crus II of cerebellar hemisphere	Right lobule IX of cerebellar hemisphere	Right precuneus
13	Left lobule III of cerebellar hemisphere	Left lobule X of cerebellar hemisphere	Right superior temporal pole
14	Right lobule III of cerebellar hemisphere	Right lobule X of cerebellar hemisphere	Right inferior temporal gyrus
15	Left lobule X of cerebellar hemisphere	Lobule I, II of vermis	Left crus II of cerebellar hemisphere
16	Lobule III of vermis	Lobule III of vermis	Left lobule III of cerebellar hemisphere
17	Lobule VIII of vermis	Lobule IV, V of vermis	Right lobule VIII of cerebellar hemisphere
18	Lobule IX of vermis	Lobule X of vermis	
19	Lobule X of vermis		

correlation methods. Besides, a suitable classifier can effectively improve classification performance and it is vital to choose a suitable one. For this reason, we perform many experiments on the multiple classifiers (e.g., LR-L1, SVM). By comparing the different classification results, we found that SVM is the most suitable one in terms of the quantitative results.

We explored the important nodes and connectivity patterns in the network of the brain constructed by the three correlation methods. In these networks, many abnormal areas of the brain and connectivity mentioned in the previous studies about OCD were found including areas in and out of the classical CSTC circuit such as the precentral gyrus and SMA (Ku et al., 2020). These results provide preliminary support for the use of the three correlation methods, especially distance correlation, as promising classification markers for patients with OCD.

Of the three FC methods, distance correlation showed the greatest diagnostic accuracy for discriminating the patients with OCD from HCS. It has been shown that distance correlation directly reflects linear and non-linear correlation in the ROIs. Therefore, the location of the regions of the brain based on distance correlation also showed a considerable research value. For example, the SMA is involved in the planning of the movement. It has been found to involve the compulsion and repetitive behavior of OCD (Gillan et al., 2016). This effect could make the distance correlation method more sensitive to detect dysfunctional neural activity than the other two FC methods. In addition, we can find that the FC matrix based on distance correlation calculation has some intergroup differences. These intergroup differences for the FC matrix may underlie the excellent classification achieved in the current study. Therefore, these ML algorithms were able to identify the patients with OCD and HCS through the FC matrix based on distance correlation calculation. This also provides support for the FC matrix composed of distance correlation calculation as a promising classification marker for OCD.

Pearson correlation was a widely used correlation method, which was generally used to measure the linear relationship between the ROIs. Therefore, its classification performance was lower than distance correlation. In addition, they showed the similarities and differences in the regions of the brain of the Pearson and distance correlation localization. These same regions of the brain had the right precentral gyrus, orbital part of the left superior frontal, and medial part of the right superior frontal gyrus, etc. In this study, they played a critical role in exploring the pathogenesis of OCD. The medial part of the right superior frontal gyrus, corresponding to the left ventral medial prefrontal cortex (vmPFC), has also been found in the disrupted emotion and cognition induced by the symptoms of OCD (Becker et al., 2014; Apergis-Schoute et al., 2017). These different regions of the brain included left putamen, right anterior cingulate, and paracingulate gyri, etc. Among them, the regions of brain (e.g., globus pellidus, putamen.) located by Pearson correlation have great research value (Hibar et al., 2018; Calzà et al., 2019). However, due to the limitation of Pearson correlation measuring linear dependency, some crucial regions of the brain will be ignored. The brain regions located (e.g., anterior cingulate and paracingulate gyri, SMA) by distance correlation can complement and extend it (Ku et al., 2020).

Partial correlation shows good classification performance, although it is lower than the Pearson and distance correlations. It was generally used to exclude the indirect influence of the correlation structure. In other words, it can measure the degree

of linear correlation between two regions of the brain without indirect influence from other regions of the brain. Therefore, we can infer that more consideration should be given to the synergies between the multiple regions of the brain in OCD. In addition, we can see that they are mainly distributed in the cerebellum in the regions of the brain defined by partial correlation. In previous studies, the cerebellum also played an important role in the exploration of the pathogenesis of OCD (Zhang et al., 2019). Therefore, we believe that partial correlation has a high potential to explore the influence of the cerebellum on OCD.

In this study, the discriminative regions of the brain are in and out of the CSTC circuit. Previous studies have reported that the orbitofrontal cortex (OFC) play crucial roles in processing reward, negative effect, and, specifically, fear and anxiety in OCD (Kringelbach and Rolls, 2004; Milad and Rauch, 2007). A recent meta-analysis of the voxel-based morphometry (VBM) studies showed decreased gray matter in the bilateral OFCs (de Wit et al., 2014). In fMRI studies, the researchers revealed the white matter abnormalities in OFC (Piras et al., 2013). Furthermore, the anterior cingulate cortex (ACC), putamen, and thalamus have been suggested to play important roles in the previous studies of OCD (Yoo et al., 2007; Zhu et al., 2015; Fan et al., 2017; Hazari et al., 2019). The OCD severity associations have been reported with hypermetabolism in the ACC (Swedo et al., 1989). In this study, consistent with these findings, the OFC, the ACC, putamen, and thalamus displayed a high degree of discriminative ability between the patients with OCD and HCS. These results provide further support for dysfunction in the CSTC circuit in patients with OCD.

In addition, some researchers found that OCD is related to the sensorimotor network (i.e., precentral gyrus/SMA) (Cui et al., 2020). Morein-Zamir et al. (2016) reported that the activation from the regions of the brain within the sensorimotor network in the inhibitory control processes may explain the essence of inhibitory control deficits of OCD. Meanwhile, one recent study indicated that OCD was associated with increased activity in the SMA. With repetitive transcranial magnetic stimulation (rTMS) treatment in SMA, the researcher can observe a reduction in the Y-BOCS score at the 4th week. The reduction in compulsion contributed to the reduction of the global Y-BOCS (Lee et al., 2017). Therefore, these previous results further support this study.

Finally, this study also showed that the cerebellum contributed to distinguish between the patients with OCD and HCS. For example, Miquel et al. (2019) suggested that inhibiting activity in the cerebellar cortex would increase impulsive and compulsive symptomatology. On the other hand, the stimulation of the cerebellar cortex should improve behavioral inhibitory control. Meanwhile, other previous studies reported the existence of disconnectedness in the fronto-striato-limbic community and connectedness between the cerebellar and visual areas in the patients with OCD, which was also related to the clinical symptomatology of OCD (Kashyap et al., 2021).

Despite the encouraging performance achieved, there are still two major limitations in the current research. First, the current study is only evaluated in a small database, which will make the results difficult to generalize. Second, the proposed method uses only single image modality data. Using a variety of modalities can obtain comprehensive features and improve the classification performance of the model. However, the subjects with the multimodal image data in the database are limited. In the future, we will explore the multimodal image data to discriminate the patients with OCD from HCS.

CONCLUSION

To conclude, the current experimental results show that it is promising to apply distance correlation for measuring the FC between the ROIs of the brain with contrast to both the traditional Pearson correlation and partial correlation. Besides improving the discrimination performance between the patients with OCD and HCS, the selected biomarkers *via* the SVM-RFE-NCV strategy may provide the potential clinical values for the patients with OCD.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Shenzhen Kangning Hospital. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

WL, ZP, and CC designed the experiment. CC obtained the funding. LJ and CC collected and preprocessed rs-fMRI data. QL performed the multivariate pattern analysis and completed the first draft of this article. WL, ZP, CC, and QL have finished the final version of this article. All the authors contributed and approved the final version of the article.

FUNDING

This study was partly supported by grants from the National Natural Science Foundation of China (61971289 and 31871113), the Shenzhen Fundamental Research Project (JCYJ20170412111316339 and JCYJ20160422113119640), and the Shenzhen-Hong Kong Institute of Brain Science—Shenzhen Fundamental Research Institutions.

ACKNOWLEDGMENTS

We sincerely thank the patients and their families for participating in the experiment and we also thank the doctors, nurses, and technicians who assisted the patients in the experiment. We also deeply appreciate the reviewers and editors for their helpful comments and suggestions for improving our work.

REFERENCES

Abramowitz, J. S., Taylor, S., and McKay, D. (2009). Obsessive-compulsive disorder. *Lancet* 374, 491–499. doi: 10.1016/S0140-6736(09)60240-3

Anderson, J. S., Nielsen, J. A., Froehlich, A. L., DuBray, M. B., Druzgal, T. J., Cariello, A. N., et al. (2011). Functional connectivity magnetic resonance imaging classification of autism. *Brain J. Neurolgy*. 134, 3742–3754. doi: 10.1093/brain/awr263

Apergis-Schoute, A. M., Gillan, C. M., Fineberg, N. A., Fernandez-Egea, E., Sahakian, B. J., and Robbins, T. W. (2017). Neural basis of impaired safety signaling in obsessive compulsive disorder. *Proc. Natl. Acad. Sci. U. S. A*. 114, 3216–3221. doi: 10.1073/pnas.1609194114

Arbabshirani, M. R., Plis, S., Sui, J., and Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *NeuroImage* 145, 137–165. doi: 10.1016/j.neuroimage.2016.02.079

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *NeuroImage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Battle, D. E. (2013). Diagnostic and statistical manual of mental disorders (DSM). *CoDAS* 25, 191–192. doi: 10.1590/s2317-17822013000200017

Becker, M. P. I., Nitsch, A. M., Schlosser, R., Koch, K., Schachtzabel, C., Wagner, G., et al. (2014). Altered emotional and BOLD responses to negative, positive and ambiguous performance feedback in OCD. *Soc. Cogn. Affect. Neur*. 9, 1127–1133. doi: 10.1093/scan/nst095

Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magnet. Reson. Med.* 34, 537–541. doi: 10.1002/mrm.1910340409

Bruin, W. B., Taylor, L., Thomas, R. M., Shock, J. P., Zhutovsky, P., Abe, Y., et al. (2020). Structural neuroimaging biomarkers for obsessive-compulsive disorder in the ENIGMA-OCD consortium: medication matters. *Transl. Psychiatr*. 10:342. doi: 10.1038/s41398-020-01013-y

Bu, X., Hu, X. Y., Zhang, L. Q., Li, B., Zhou, M., Lu, L., et al. (2019). Investigating the predictive value of different resting-state functional MRI parameters in obsessive-compulsive disorder. *Transl. Psychiat*. 9:17. doi: 10.1038/s41398-018-0362-9

Calzà, J., Gürsel, D. A., Schmitz-Koep, B., Bremer, B., Reinholz, L., Berberich, G., et al. (2019). Altered cortico-striatal functional connectivity during resting state in obsessive-compulsive disorder. *Front. psychiat*. 10:319. doi: 10.3389/fpsyt.2019.00319

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA). doi: 10.1145/2939672.2939785

Cui, G., Ou, Y., Chen, Y., Lv, D., Jia, C., Zhong, Z., et al. (2020). Altered global brain functional connectivity in drug-naive patients with obsessive-compulsive disorder. *Front. Psychiatr*. 11:98. doi: 10.3389/fpsyt.2020.00098

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., et al. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage* 192, 115–134. doi: 10.1016/j.neuroimage.2019.02.062

de Vries, F. E., de Wit, S. J., van den Heuvel, O. A., Veltman, D. J., Cath, D. C., van Balkom, A. J. L. M., et al. (2019). Cognitive control networks in OCD: a resting-state connectivity study in unmedicated patients with obsessive-compulsive disorder and their unaffected relatives. *World J. Biol. Psychia*. 20, 230–242. doi: 10.1080/15622975.2017.1353132

de Wit, S. J., Alonso, P., Schweren, L., Mataix-Cols, D., Lochner, C., Menchón, J. M., et al. (2014). Multicenter voxel-based morphometry mega-analysis of structural brain scans in obsessive-compulsive disorder. *Am. J. Psychiatr*. 171, 340–349. doi: 10.1176/appi.ajp.2013.13040574

Ding, X., Yang, Y., Stein, E. A., and Ross, T. J. (2015). Multivariate classification of smokers and nonsmokers using SVM-RFE on structural MRI images. *Hum. Brain Map*. 36, 4869–4879. doi: 10.1002/hbm.22956

Fan, S., Cath, D. C., van den Heuvel, O. A., van der Werf, Y. D., Schöls, C., Veltman, D. J., et al. (2017). Abnormalities in metabolite concentrations in tourette's disorder and obsessive-compulsive disorder-a proton magnetic resonance spectroscopy study. *Psychoneuroendocrinology* 77, 211–217. doi: 10.1016/j.psyneuen.2016.12.007

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10, 507–521. doi: 10.2307/2331838

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning: Springer Series in Statistics*. New York, NY: Springer.

Gillan, C. M., Robbins, T. W., Sahakian, B. J., van den Heuvel, O. A., and van Wingen, G. (2016). The role of habit in compulsivity. *Eur. Neuropsychopharm*. 26, 828–840. doi: 10.1016/j.euroneuro.2015.12.033

Goodman, W. K., Grice, D. E., Lapidus, K. A., and Coffey, B. J. (2014). Obsessive-compulsive disorder. *Psychiatr. Clin. North Am*. 37, 257–267. doi: 10.1016/j.psc.2014.06.004

Gruner, P., Vo, A., Argyelan, M., Ikuta, T., Degnan, A. J., John, M., et al. (2014). Independent component analysis of resting state activity in pediatric obsessive-compulsive disorder. *Hum. Brain Map*. 35, 5306–5315. doi: 10.1002/hbm.22551

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learn.* 46, 389–422. doi: 10.1023/A:1012487302797

Hazari, N., Narayanaswamy, J. C., and Venkatasubramanian, G. (2019). Neuroimaging findings in obsessive-compulsive disorder: a narrative review to elucidate neurobiological underpinnings. *Ind. J. Psychiat*. 61, S9–S29. doi: 10.4103/psychiatry.IndianJPsychiatry_525_18

Hibar, D. P., Cheung, J. W., Medland, S. E., Mufford, M. S., Jahanshad, N., Dalvie, S., et al. (2018). Significant concordance of genetic variation that increases both the risk for obsessive-compulsive disorder and the volumes of the nucleus accumbens and putamen. *British J. Psychiat.* 213, 430–436. doi: 10.1192/bjp.2018.62

Hou, J. M., Zhao, M., Zhang, W., Song, L. H., Wu, W. J., Wang, J., et al. (2014). Resting-state functional connectivity abnormalities in patients with obsessive-compulsive disorder and their healthy first-degree relatives. *J. Psychiatr. Neurosci*. 39, 304–311. doi: 10.1503/jpn.130220

Kashyap, R., Eng, G. K., Bhattacharjee, S., Gupta, B., Ho, R., Ho, C. S. H., et al. (2021). Individual-fMRI-approaches reveal cerebellum and visual communities to be functionally connected in obsessive compulsive disorder. *Sci. Rep*. 11:1354. doi: 10.1038/s41598-020-80346-6

Kassraian-Fard, P., Matthis, C., Balsters, J. H., Maathuis, M. H., and Wenderoth, N. (2016). Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Front. Psychiatry* 7:177. doi: 10.3389/fpsyt.2016.00177

Kipf, T. N. (2017). *Semi-Supervised Classification With Graph Convolutional Networks*. Toulon.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain J. Neurol*. 131, 681–689. doi: 10.1093/brain/awm319

Kringelbach, M. L., and Rolls, E. T. (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progr. Neurobiol*. 72, 341–372. doi: 10.1016/j.pneurobio.2004.03.006

Ku, J., Kim, S. J., Lee, H., Jhung, K., An, S. K., Namkoong, K., et al. (2020). Deactivation of anterior cingulate cortex during virtual social interaction in obsessive-compulsive disorder. *Psychiatr. Res. Neuroimag.* 304:111154. doi: 10.1016/j.pscychresns.2020.111154

Lamothe, H., Baleyte, J. M., Smith, P., Pelissolo, A., and Mallet, L. (2018). Individualized immunological data for precise classification of OCD patients. *Brain Sci*. 8:80149. doi: 10.3390/brainsci8080149

Lee, Y. J., Koo, B. H., Seo, W. S., Kim, H. G., Kim, J. Y., and Cheon, E. J. (2017). Repetitive transcranial magnetic stimulation of the supplementary motor area in treatment-resistant obsessive-compulsive disorder: an open-label pilot study. *J. Clin. Neurosci*. 44, 264–268. doi: 10.1016/j.jocn.2017.06.057

Lin, X., Li, C., Zhang, Y., Su, B., Fan, M., and Wei, H. (2017). Selecting feature subsets based on SVM-RFE and the overlapping ratio with applications in bioinformatics. *Molecules*. 23:10052. doi: 10.3390/molecules23010052

Liu, F., Guo, W., Fouche, J. P., Wang, Y., Wang, W., Ding, J., et al. (2015a). Multivariate classification of social anxiety disorder using whole brain functional connectivity. *Brain Struct. Funct*. 220, 101–115. doi: 10.1007/s00429-013-0641-4

Liu, F., Wang, Y., Li, M., Wang, W., Li, R., Zhang, Z., et al. (2017). Dynamic functional network connectivity in idiopathic generalized epilepsy

with generalized tonic-clonic seizure. *Hum. Brain Map*. 38, 957–973. doi: 10.1002/hbm.23430

Liu, F., Wee, C. Y., Chen, H., and Shen, D. (2014). Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification. *NeuroImage* 84, 466–475. doi: 10.1016/j.neuroimage.2013.09.015

Liu, F., Xie, B., Wang, Y., Guo, W., Fouche, J. P., Long, Z., et al. (2015b). Characterization of post-traumatic stress disorder using resting-state fMRI with a multi-level parametric classification approach. *Brain Topogr*. 28, 221–237. doi: 10.1007/s10548-014-0386-2

Menzies, L., Chamberlain, S. R., Laird, A. R., Thelen, S. M., Sahakian, B. J., and Bullmore, E. T. (2008b). Integrating evidence from neuroimaging and neuropsychological studies of obsessive-compulsive disorder: the orbitofronto-striatal model revisited. *Neurosci. Biobehav. Rev*. 32, 525–549. doi: 10.1016/j.neubiorev.2007.09.005

Menzies, L., Williams, G. B., Chamberlain, S. R., Ooi, C., Fineberg, N., Suckling, J., et al. (2008a). White matter abnormalities in patients with obsessive-compulsive disorder and their first-degree relatives. *Am. J. Psychiatr*. 165, 1308–1315. doi: 10.1176/appi.ajp.2008.07101677

Milad, M. R., and Rauch, S. L. (2007). The role of the orbitofrontal cortex in anxiety disorders. *Ann. N. Y. Acad. Sci*. 1121, 546–561. doi: 10.1196/annals.1401.006

Miquel, M., Nicola, S. M., Gil-Miravet, I., Guarque-Chabrera, J., and Sanchez-Hernandez, A. (2019). A working hypothesis for the role of the cerebellum in impulsivity and compulsivity. *Front. Behav. Neurosci*. 13:99. doi: 10.3389/fnbeh.2019.00099

Morein-Zamir, S., Voon, V., Dodds, C. M., Sule, A., van Niekerk, J., Sahakian, B. J., et al. (2016). Divergent subcortical activity for distinct executive functions: stopping and shifting in obsessive compulsive disorder. *Psychol. Med*. 46, 829–840. doi: 10.1017/S0033291715002330

Mueller, S., Wang, D., Pan, R., Holt, D. J., and Liu, H. (2015). Abnormalities in hemispheric specialization of caudate nucleus connectivity in schizophrenia. *J. Am. Med. Assoc. Psychiatr*. 72, 552–560. doi: 10.1001/jamapsychiatry.2014.3176

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci. Biobehav. Rev*. 36, 1140–1152. doi: 10.1016/j.neubiorev.2012.01.004

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2012). Scikit-learn: machine learning in python. *J. Mac. Learn. Res*. 12, 2825–2830.

Piacentini, J., Bergman, R. L., Keller, M., and McCracken, J. (2003). Functional impairment in children and adolescents with obsessive-compulsive disorder. *J. Child Adolesc. Psychopharmacol*. 13(Suppl.1), S61–S69. doi: 10.1089/104454603322126359

Piras, F., Piras, F., Caltagirone, C., and Spalletta, G. (2013). Brain circuitries of obsessive compulsive disorder: a systematic review and meta-analysis of diffusion tensor imaging studies. *Neurosci. Biobehav. Rev*. 37, 2856–2877. doi: 10.1016/j.neubiorev.2013.10.008

Plitt, M., Barnes, K. A., Wallace, G. L., Kenworthy, L., and Martin, A. (2015). Resting-state functional connectivity predicts longitudinal change in autistic traits and adaptive functioning in autism. *Proc. Natl. Acad. Sci. U. S. A*. 112, E6699–E6706. doi: 10.1073/pnas.1510098112

Rapinesi, C., Kotzalidis, G. D., Ferracuti, S., Sani, G., Girardi, P., and Del Casale, A. (2019). Brain stimulation in obsessive-compulsive disorder (OCD): a systematic review. *Curr. Neuropharmacol*. 17, 787–807. doi: 10.2174/1570159X17666190409142555

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155, 530–548. doi: 10.1016/j.neuroimage.2017.03.057

Rehn, S., Eslick, G. D., and Brakoulias, V. (2018). A meta-analysis of the effectiveness of different cortical targets used in repetitive transcranial magnetic stimulation (rTMS) for the treatment of obsessive-compulsive disorder (OCD). *Psychiatr. Quart*. 89, 645–665. doi: 10.1007/s11126-018-9566-7

Ruscio, A. M., Stein, D. J., Chiu, W. T., and Kessler, R. C. (2010). The epidemiology of obsessive-compulsive disorder in the National Comorbidity Survey Replication. *Mol. Psychiatr*. 15, 53–63. doi: 10.1038/mp.2008.94

Sajda, P. (2006). Machine learning for detection and diagnosis of disease. *Ann. Rev. Biomed. Eng*. 8, 537–565. doi: 10.1146/annurev.bioeng.8.061505.095802

Saxena, S., Brody, A. L., Schwartz, J. M., and Baxter, L. R. (1998). Neuroimaging and frontal-subcortical circuitry in obsessive-compulsive disorder. *Br. J. Psychiat*. 173, 26–37. doi: 10.1192/S0007125000297870

Sen, B., Bernstein, G. A., Tingting, X., Mueller, B. A., Schreiner, M. W., Cullen, K. R., et al. (2016). "Classification of obsessive-compulsive disorder from resting-state fMRI," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Orlando), 3606–3609. doi: 10.1109/EMBC.2016.7591508

Shenas, S. K., Halici, U., and Cicek, M. (2013). "Detection of obsessive compulsive disorder using resting-state functional connectivity data," in *Proceedings of the 2013 6th International Conference on Biomedical Engineering and Informatics*, eds J. X. Gao, D. Xu, X. Sun, and Y. Wu (Chicago, IL), 132–136. doi: 10.1109/BMEI.2013.6746921

Shenas, S. K., Halici, U., and Çiçek, M. (2014). "A comparative analysis of functional connectivity data in resting and task-related conditions of the brain for disease signature of OCD," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Hangzhou), 978–981. doi: 10.1109/EMBC.2014.6943756

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., et al. (2011). Network modelling methods for FMRI. *NeuroImage* 54, 875–891. doi: 10.1016/j.neuroimage.2010.08.063

Stein, D. J., Kogan, C. S., Atmaca, M., Fineberg, N. A., Fontenelle, L. F., Grant, J. E., et al. (2016). The classification of obsessive-compulsive and related disorders in the ICD-11. *J. Affect. Disord*. 190, 663–674. doi: 10.1016/j.jad.2015.10.061

Swedo, S. E., Schapiro, M. B., Grady, C. L., Cheslow, D. L., Leonard, H. L., Kumar, A., et al. (1989). Cerebral glucose metabolism in childhood-onset obsessive-compulsive disorder. *Arch. Gen. Psychiatr*. 46, 518–523. doi: 10.1001/archpsyc.1989.01810060038007

Szekely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Stat*. 35, 2769–2794. doi: 10.1214/009053607000000505

Takagi, Y., Sakai, Y., Lisi, G., Yahata, N., Abe, Y., Nishida, S., et al. (2017). A neural marker of obsessive-compulsive disorder from whole-brain functional connectivity. *Sci. Rep*. 7:7538. doi: 10.1038/s41598-017-07792-7

Thorsen, A. L., Hagland, P., Radua, J., Mataix-Cols, D., Kvale, G., Hansen, B., et al. (2018). Emotional processing in obsessive-compulsive disorder: a systematic review and meta-analysis of 25 functional neuroimaging studies. *Biol. Psychiatr. Cogn. Neurosci. Neuroimag*. 3, 563–571. doi: 10.1016/j.bpsc.2018.01.009

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Varoquaux, G., Gramfort, A., Poline, J. B., and Thirion, B. (2010). *Brain Covariance Selection: Better Individual Functional Connectivity Models Using Population Prior. Advances in Neural Information Processing Systems; 2010 2010-12-07*; Vancouver, Canada. Available online at: https://hal.inria.fr/inria-00512451v4/document; https://hal.inria.fr/inria-00512451v4/file/paper.pdf (accessed September 27, 2021).

Vergun, S., Deshpande, A., Meier, T., Song, J., Tudorascu, D., Nair, V., et al. (2013). Characterizing functional connectivity differences in aging adults using machine learning on resting state fMRI data. *Front. Comput. Neurosci*. 7:38. doi: 10.3389/fncom.2013.00038

Wang, C., Xiao, Z., and Wu, J. (2019). Functional connectivity-based classification of autism and control using SVM-RFECV on rs-fMRI data. *Phys. Med*. 65, 99–105. doi: 10.1016/j.ejmp.2019.08.010

Xia, M., Wang, J., and He, Y. (2013). BrainNet Viewer: a network visualization tool for human brain connectomics. *PLoS ONE* 8:e68910. doi: 10.1371/journal.pone.0068910

Xue, K., Liang, S., Yang, B., Zhu, D., Xie, Y., Qin, W., et al. (2020). Local dynamic spontaneous brain activity changes in first-episode, treatment-naïve patients with major depressive disorder and their associated gene expression profiles. *Psychol. Med*. 2020, 1–10. doi: 10.1017/S0033291720003876

Yang, X., Luo, J., Zhong, Z., Yang, X., Yao, S., Wang, P., et al. (2019). Abnormal regional homogeneity in patients with obsessive-compulsive disorder and

their unaffected siblings: a resting-state fMRI study. *Front. Psychiatr*. 10:452. doi: 10.3389/fpsyt.2019.00452

Yoo, K., Rosenberg, M. D., Noble, S., Scheinost, D., Constable, R. T., and Chun, M. M. (2019). Multivariate approaches improve the reliability and validity of functional connectivity and prediction of individual behaviors. *NeuroImage* 197, 212–223. doi: 10.1016/j.neuroimage.2019.04.060

Yoo, S. Y., Jang, J. H., Shin, Y. W., Kim, D. J., Park, H. J., Moon, W. J., et al. (2007). White matter abnormalities in drug-naïve patients with obsessive-compulsive disorder: a diffusion tensor study before and after citalopram treatment. *Acta Psychiatr. Scand*. 116, 211–219. doi: 10.1111/j.1600-0447.2007.01046.x

Zeng, L. L., Shen, H., Liu, L., Wang, L., Li, B., Fang, P., et al. (2012). Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. *Brain J. Neurol*. 135, 1498–1507. doi: 10.1093/brain/aws059

Zhan, Y., Wei, J., Liang, J., Xu, X., He, R., Robbins, T. W., et al. (2021). Diagnostic classification for human autism and obsessive-compulsive disorder based on machine learning from a primate genemodel. *Am. J. Psychiatr*. 178, 65–76. doi: 10.1176/appi.ajp.2020.19101091

Zhang, H., Wang, B., Li, K., Wang, X., Li, X., Zhu, J., et al. (2019). Altered functional connectivity between the cerebellum and the cortico-striato-thalamo-cortical circuit in obsessive-compulsive disorder. *Front. Psychiatr*. 10:522. doi: 10.3389/fpsyt.2019.00522

Zhou, C., Cheng, Y., Ping, L., Xu, J., Shen, Z., Jiang, L., et al. (2018). Support vector machine classification of obsessive-compulsive disorder based on whole-brain volumetry and diffusion tensor imaging. *Front. Psychiatr*. 9:524. doi: 10.3389/fpsyt.2018.00524

Zhu, Y., Fan, Q., Han, X., Zhang, H., Chen, J., Wang, Z., et al. (2015). Decreased thalamic glutamate level in unmedicated adult obsessive-compulsive disorder patients detected by proton magnetic resonance spectroscopy. *J. Affect. Disord*. 178, 193–200. doi: 10.1016/j.jad.2015.03.008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 19 November 2021
doi: 10.3389/fninf.2021.715421

Comparison of Resampling Techniques for Imbalanced Datasets in Machine Learning: Application to Epileptogenic Zone Localization From Interictal Intracranial EEG Recordings in Patients With Focal Epilepsy

Giulia Varotto[1,2]*, *Gianluca Susi*[3,4], *Laura Tassi*[5], *Francesca Gozzo*[5], *Silvana Franceschetti*[2] and *Ferruccio Panzica*[6]

[1] Epilepsy Unit, Bioengineering Group, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy, [2] Neurophysiopathology Unit, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy, [3] Universidad Complutense de Madrid-Universidad Politécnica de Madrid (UPM-UCM) Laboratory of Cognitive and Computational Neuroscience, Center of Biomedical Technology, Technical University of Madrid, Madrid, Spain, [4] Department of Experimental Psychology, Cognitive Processes and Logopedy, Complutense University of Madrid, Madrid, Spain, [5] "Claudio Munari" Epilepsy Surgery Centre, Niguarda Hospital, Milan, Italy, [6] Clinical Engineering, Fondazione IRCCS Istituto Neurologico Carlo Besta, Milan, Italy

OPEN ACCESS

Edited by:
Sharlene D. Newman, University of Alabama, United States

Reviewed by:
Leon D. Iasemidis, Louisiana Tech University, United States
Joseph R. Madsen, Boston Pediatric Neurosurgical Foundation, United States

***Correspondence:**
Giulia Varotto
giulia.varotto@istituto-besta.it
orcid.org/0000-0001-8849-0398

Received: *26 May 2021*
Accepted: *23 September 2021*
Published: *19 November 2021*

Citation:
Varotto G, Susi G, Tassi L, Gozzo F, Franceschetti S and Panzica F (2021) Comparison of Resampling Techniques for Imbalanced Datasets in Machine Learning: Application to Epileptogenic Zone Localization From Interictal Intracranial EEG Recordings in Patients With Focal Epilepsy. Front. Neuroinform. 15:715421. doi: 10.3389/fninf.2021.715421

Aim: In neuroscience research, data are quite often characterized by an imbalanced distribution between the majority and minority classes, an issue that can limit or even worsen the prediction performance of machine learning methods. Different resampling procedures have been developed to face this problem and a lot of work has been done in comparing their effectiveness in different scenarios. Notably, the robustness of such techniques has been tested among a wide variety of different datasets, without considering the performance of each specific dataset. In this study, we compare the performances of different resampling procedures for the imbalanced domain in stereo-electroencephalography (SEEG) recordings of the patients with focal epilepsies who underwent surgery.

Methods: We considered data obtained by network analysis of interictal SEEG recorded from 10 patients with drug-resistant focal epilepsies, for a supervised classification problem aimed at distinguishing between the epileptogenic and non-epileptogenic brain regions in interictal conditions. We investigated the effectiveness of five oversampling and five undersampling procedures, using 10 different machine learning classifiers. Moreover, six specific ensemble methods for the imbalanced domain were also tested. To compare the performances, Area under the ROC curve (AUC), F-measure, Geometric Mean, and Balanced Accuracy were considered.

Results: Both the resampling procedures showed improved performances with respect to the original dataset. The oversampling procedure was found to be more sensitive to the type of classification method employed, with Adaptive Synthetic Sampling

(ADASYN) exhibiting the best performances. All the undersampling approaches were more robust than the oversampling among the different classifiers, with Random Undersampling (RUS) exhibiting the best performance despite being the simplest and most basic classification method.

Conclusions: The application of machine learning techniques that take into consideration the balance of features by resampling is beneficial and leads to more accurate localization of the epileptogenic zone from interictal periods. In addition, our results highlight the importance of the type of classification method that must be used together with the resampling to maximize the benefit to the outcome.

Keywords: imbalanced dataset classification, re-sampling techniques, oversampling and undersampling, ensemble methods, network analysis, epilepsy surgery, stereo-EEG/intracranial recordings, epileptogenic zone localization

INTRODUCTION

Epilepsy is a chronic neurological disease affecting 1% of the worldwide population (Fiest et al., 2017). Approximately 30% of the patients with focal epilepsies are resistant to the antiepileptic drugs (AEDs), and they can be considered as candidate for epilepsy surgery, with the aim of removing the epileptogenic zone (EZ). The latter is defined as the minimum amount of cortex that must be resected (inactivated or completely disconnected) to produce seizure freedom (Lüders et al., 2006; Ryvlin et al., 2014). However, the correct localization of the EZ to achieve seizure freedom after surgery, is still an unsolved and open question, as indicated by the high rate of failure of seizure control (30–40%) after surgery (Spencer and Huh, 2008; Bulacio et al., 2012). The advanced signal processing approaches, especially those based on the connectivity analysis, have been largely applied to stereo-electroencephalography (SEEG) from the patients with epilepsy to better pinpoint the location of the EZ (Varotto et al., 2013; Bartolomei et al., 2017; Adkinson et al., 2019; Narasimhan et al., 2020).

The supervised machine learning methods are increasingly applied in epilepsy research, representing useful tools to integrate the complex and large-scale data deriving from different electrophysiological or imaging techniques, such as EEG, magnetoencephalography (MEG), functional-MRI (fMRI), or positron emission tomography (PET) (refer to Abbasi and Goldenholz, 2019 for a comprehensive review). Most of these studies focused on the following aspects: diagnosis of epilepsy (Kassahun et al., 2014; Azami et al., 2016; Soriano et al., 2017), seizure prediction (Acharya et al., 2018; Kiral-Kornek et al., 2018; Daoud and Bayoumi, 2019), lateralization of temporal lobe epilepsy (Jin and Chung, 2017; Frank et al., 2018; Peter et al., 2018), and post-surgical outcome prediction (Armañanzas et al., 2013; Goldenholz et al., 2016; Gleichgerrcht et al., 2018). With respect to the localization of the EZ and support to pre-surgical planning, few works applied machine learning tools, showing the promising usefulness of this approach, and the need for further investigation and generalization (Dian et al., 2015; Elahian et al., 2017; Khambhati et al., 2017; Roland et al., 2017). In this specific framework, one central issue that should be taken into account, and which could represent one of the main limitations, is that the EZ represents a smaller region compared with the other non-EZ areas explored. This leads to an uneven distribution of the majority (non-EZ) and minority (EZ) classes, which can strongly worsen or limit the classification performances. This situation is known as the class imbalance problem and can be considered one of the central topics in machine learning research (He and Garcia, 2009; Ali et al., 2015; Fernández et al., 2018).

In the past decade, many different approaches have been developed to cope with imbalanced classification, most of them based on four different families: resampling techniques, cost-sensitive learning, algorithm modification, and ensemble methods (Mena and Gonzalez, 2006; Galar et al., 2012; Krawczyk et al., 2014; Loyola-González et al., 2016).

Among these, the methods belonging to the data resampling family have been proved useful as well as relatively simple approaches to be applied in the medical context (Lee, 2014; Loyola-González et al., 2016). In data resampling, the training instances are modified to rebalance the class distribution through *oversampling* of the minority class, or *undersampling* of the majority one, before training the classifier. Oversampling could have the limitation of overfitting the minority class, while undersampling could eliminate potential useful information for correct classification (Chawla, 2009).

Different studies dealt with the comparisons of performances of most of the existing resampling techniques, most of which were applied to a wide variety of datasets together, being mainly aimed at assessing the robustness of results across different dataset combinations (López et al., 2013). Nevertheless, when applied to a single specific dataset, such comparison can lead to different results (Xie et al., 2020), reflecting a lack of consensus about the performances of such techniques and putting in evidence the need for *ad-hoc* comparisons in each specific clinical framework.

To the best of our knowledge, this is the first study focused on the evaluation and comparison of these approaches in the context of epilepsy, and in particular, in the framework of the surgical planning based on analysis of electrophysiological intracranial recordings.

In this study, we compared five oversampling and five undersampling procedures and tested the resulting rebalanced datasets with 10 different machine learning classifiers (such as

both standard machines and classical ensemble approaches). Moreover, six specific ensemble methods properly modified for imbalanced domain and belonging to data variation-based ensemble were tested and compared. In these algorithms, the resampling phase is applied to each step of the ensemble classifier, in such a way that each classifier is trained with a different resampled dataset (Galar et al., 2012). For this reason, we considered them as an extension of resampling methods, which need to be compared with the oversampling and undersampling techniques combined with the classical ensemble approaches.

The classification was based on the features obtained by network analysis of interictal SEEG recorded from the 10 patients who underwent epilepsy surgery and were seizure-free (SF) after 3 years of follow-up.

To compare the performances, *area under the ROC curve* (AUC), *balanced accuracy* (BalACC), *F-measure* (Fm), and *geometric mean* (Gmean) were used as metrics, since these are usually considered suitable measures to deal with the imbalanced datasets (Bekkar et al., 2013; López et al., 2013).

MATERIALS AND METHODS

We start this section by describing the steps of selection and signal recording of the patients. The methodological pipeline is then outlined: feature extraction, data resampling, classification, and evaluation of the performance of the model (as shown in **Figure 1** for a schematic representation). Finally, we describe the statistical analysis, which has been performed to evaluate the consistency of our results.

Selection of Patients

The study involved SEEG signals recorded from $Np = 10$ patients (three women) with drug-resistant focal epilepsy at the Claudio Munari Epilepsy Surgery Center of Niguarda Hospital (Milan, Italy). The patients were selected from the 41 patients implanted with SEEG electrodes over 24 months. Among them, 24 had negative MRI and 10 of them were seizure-free after at least 3 years of follow-up and were finally considered for this study. **Table 1** presents the details of the main clinical features.

The mean age of the patients was 31.7 $\pm$ 7.3 years, and the mean duration of epilepsy was 17.2 $\pm$ 7.8 years. They had no obvious risk factor for epilepsy. The surgical outcome was assessed after at least 3 years of follow-up after surgery (mean follow-up period: 56 $\pm$ 13 months) and classified as class I according to Engel's classification (Engel, 1993).

SEEG Recordings

Stereo-electroencephalography signals were recorded using the multi-lead platinum-iridium electrodes (Dixi, Besançon, France, with 5–18 contacts of diameter 0.8 mm; 1.5 mm long; and 2 mm apart), implanted under general anesthesia after stereo-arteriography using a 3D MRI imported into a computer-assisted neuronavigational module to localize the blood vessels and guide electrode trajectory. The placement of intracerebral electrodes was defined according to the data derived by non-invasive anatomo-electroclinical procedures (Talairach and Bancaud, 1966; Cardinale et al., 2019).

The SEEG signals were recorded using a common reference electrode (Nikon-Kohden system; 192-channels; sampling rate 1 kHz) under video and clinical control over 5–20 days and then examined by the two expert neurologists to define the EZ and plan the surgical approach and resection. EZ was defined by considering ictal discharge recordings, responses associated with the intracerebral electrical stimulations, and neurophysiological mapping, and then integrated into the definition of the brain area(s) to be surgically excised. Post-resection MRI was used to identify the areas of the brain that were effectively removed. The target value to assess the classification performances—SEEG leads as belonging to EZ or non-EZ—was defined by considering the intersection between the group of SEEG leads labeled as EZ by the clinicians through the pre-surgical evaluation, and the resected zone.

Feature Extraction

Stereo-electroencephalography signals were analyzed using bipolar derivations, and those presenting non-physiological artifacts were excluded from the analysis. The number of analyzed SEEG leads differed for each patient being on average $N_L = 73 \pm 6$. Furthermore, 3 min of continuous interictal SEEG signals, recorded during awake condition at least 1 h far from any ictal event, were selected and divided into $N_E = 36$, five s length, non-overlapping epochs. After testing several lengths and epochs partitions, 3 min length was selected as the minimum recording time to obtain a good EZ classification. The broad 1–80 Hz frequency band was used for the analysis. In addition, 36 time-varying connectivity matrices were estimated by applying a bivariate non-linear method and the non-linear regression index (h2) (Lopes da Silva et al., 1989; Wendling et al., 2010) (refer to **Supplementary Material**). In this regard, a wide variety of methods have been proposed to estimate the SEEG connectivity, all of them being characterized by different advantages and pitfalls strongly depending on the signal and the aim of the study (Silfverhuth et al., 2012; Olejarczyk et al., 2017). Among them, a non-linear regression analysis has been proved to be particularly suitable to estimate the connectivity from the simulated coupled neuronal population (Wendling et al., 2009), and has been largely applied in the specific contest of intracranial EEG recordings and EZ localization (Bartolomei et al., 2017).

From the adjacency matrices, the corresponding graphs were built for each patient, after applying a threshold to select the minimum number of connections that ensures a connected graph for all the epochs.

After a preliminary analysis involving several graph theory-based indices, nine of them, focusing on different complementary network properties of centrality (Oldham et al., 2019), were identified as the optimal one to classify EZ in the whole group of patients, and used as features of the classifier: *outdegree centrality (Ce), indegree Ce, oustrength Ce, instrength Ce., betweenness Ce., outcloseness Ce., incloseness Ce., pagerank Ce.,* and *eigenvector Ce.* (as shown in **Supplementary Material** for a detailed description of the basic properties of these metrics).

The connectivity analysis was performed through a specific custom-written toolbox developed in Matlab (R20a; MathWorks Inc., Natick, MA, USA). Matlab graph toolbox and the Brain

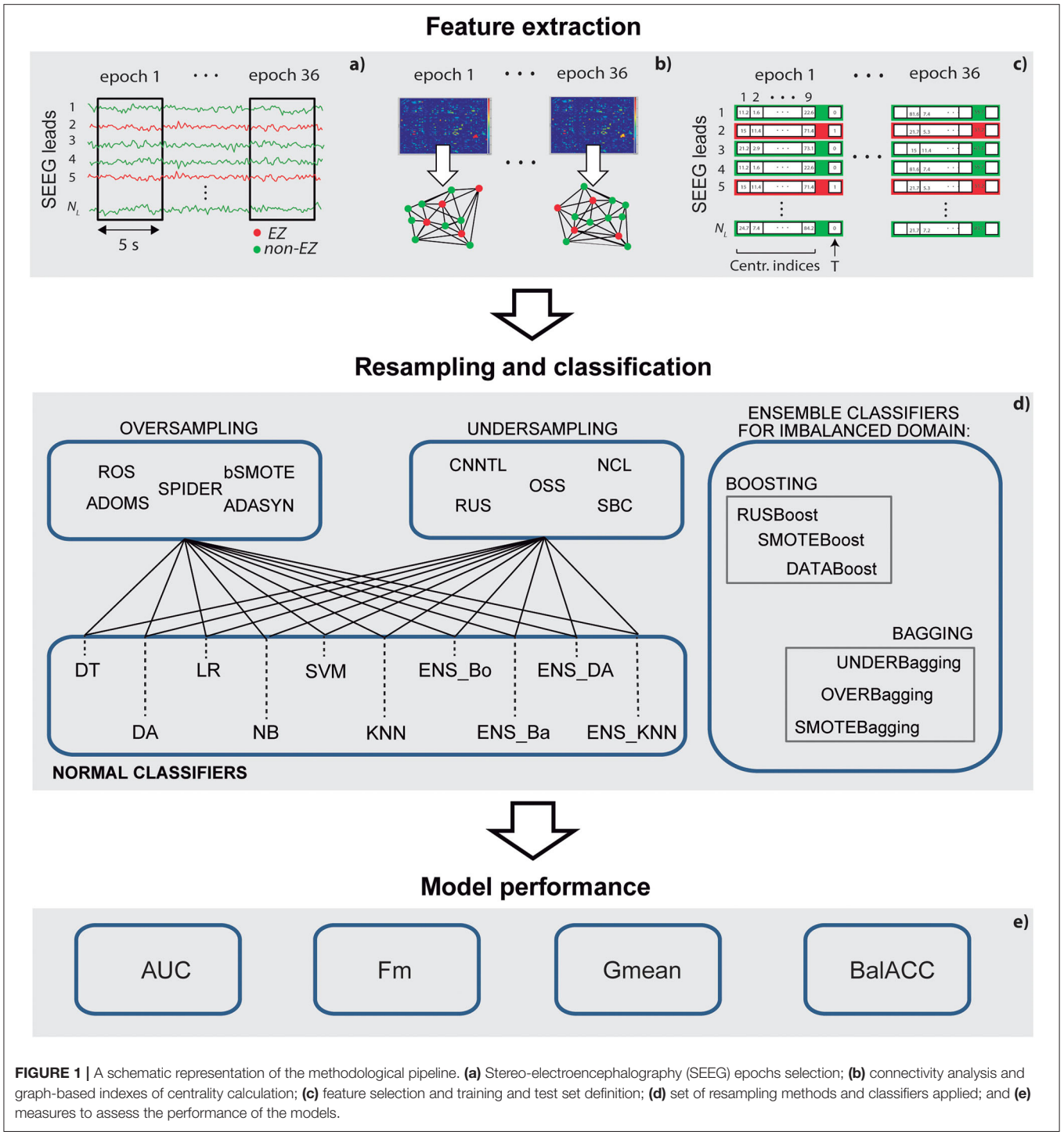


FIGURE 1 | A schematic representation of the methodological pipeline. **(a)** Stereo-electroencephalography (SEEG) epochs selection; **(b)** connectivity analysis and graph-based indexes of centrality calculation; **(c)** feature selection and training and test set definition; **(d)** set of resampling methods and classifiers applied; and **(e)** measures to assess the performance of the models.

connectivity toolbox (Rubinov and Sporns, 2010), were used for graph analysis.

To provide the classifiers with a suitable number of trials, we first grouped all the values of the features pertaining to the different time epochs and obtained, for each patient p, a matrix with $N_{L,p} \times N_E$ rows and 10 columns (i.e., nine features and one target). EZ has been considered as the positive class, with 1 indicating the EZ class and 0 the non-EZ class. The imbalanced ratio (IR)—the ratio between the number of trials pertaining to positive and negative classes—for each patient, is indicated in **Table 2**.

Since one of the main objectives of the proposed procedure was to classify SEEG signals of every single patient independently from the others, training and test set were defined by considering

TABLE 1 | Main clinical features, epileptogenic zone (EZ) localization performed by the standard methods, and surgery outcomes for the patients enrolled in this study.

Id	Gender	Age	Onset	Sz/m	Side	Lobe	Histology	Follow-Up/M	AEDs
1	M	27	19	10	R	TFI	crypto	68	Reduced
2	M	25	4	10	L	TI	crypto	48	Stopped
3	F	30	16	5	R	T	crypto	54	Reduced
4	F	27	17	10	R	T	crypto	46	Stopped
5	M	40	16	30	L	F	no	70	Reduced
6	M	39	20	1	R	F	crypto	42	Reduced
7	M	28	11	3	L	F	crypto	65	Ongoing
8	M	22	16	15	L	TO	crypto	34	Reduced
9	M	44	22	10	R	TO	FCD Ib	62	Reduced
10	F	35	4	5	R	TPCF	FCD Ia	71	Ongoing

AEDs, antepileptic drugs; crypto, cryptogenic; FCD, focal cortical dysplasia; F, female; FC, fronto-central; FCD, focal cortical dysplasia; Fr, Frontal; HS, hippocampal sclerosis; M, male; Sz//m, seizures per month; Age, age at surgical intervention; Onset, age of epilepsy onset; PCI, parieto-centro-insular; T, Temporal; TFI, temporo-fronto-insular; TI, temporo insular; TC, temporo-central; TO, temporo-occipital; TPCF, temporo-parieto-centro-frontal. AEDs column refers to variation of drug therapy with respect to pre-surgical condition.

TABLE 2 | Number of analyzed SEEG leads (Total SEEG leads), number of leads belonging to the EZ (EZ leads), and Imbalanced ratio (IR) per patient.

Pt id	Tot SEEG leads	EZ leads	IR
1	66	5	12.2
2	73	6	11.2
3	80	7	10.4
4	81	9	8.0
5	62	4	14.5
6	72	2	35.0
7	76	8	8.5
8	78	13	5.0
9	72	5	13.4
10	72	7	9.3

a proportion of 9:1, using features from nine patients for training and features from one single patient for test. For further statistical analysis, the same splitting was repeated for all the combinations of patients, thus providing 10 different training-testing datasets.

Data Resampling

In all the patients, more electrode contacts were implanted in the non-epileptogenic than epileptogenic regions. This fact is reflected in a smaller number of EZ trials than the non-EZ trials, giving rise to the problems with the statistics of the applied classification methods (and hence, the subsequent learning by machine learning models).

Among the existing resampling techniques to tackle such class imbalance problems, we selected five methods of oversampling and five methods of undersampling and compared the performance of classifiers with respect to the original dataset.

The oversampling methods are based on the creation of a new bigger dataset, obtained by replicating or creating new samples, usually from the minority class:

- *Adaptive Synthetic Sampling* (ADASYN). ADASYN generates data considering a weighted distribution for different minority class examples, where more synthetic data are generated for minority class examples that are harder to learn compared with those easier to learn (He et al., 2008).
- *Adjusting the direction of the synthetic minority class example* (ADOMS). ADOMS generates positive data instances from other instances in the original dataset selecting k as the nearest neighbors and using them to perform arithmetical operations to generate the new instance by principal component analysis (PCA) (Tang and Chen, 2008).
- *Random oversampling* (ROS). ROS generates minority class instances randomly (Batista et al., 2004).
- Selective Pre-processing for Imbalanced Data (SPIDER). SPIDER oversamples instances from the minority class that are difficult to learn and, at the same time, filters the examples from the majority class which are also difficult to learn (Stefanowski and Wilk, 2008).
- *Borderline-Synthetic Monitoring Oversampling Technique* (bSMOTE). The bSMOTE generates positive data instances from other instances in the original dataset selecting k as the nearest neighbors and using them to perform the arithmetical operations to generate the new instance (Han et al., 2005).

The undersampling methods are based on the reduction of the original dataset by eliminating samples, usually form the majority class:

- Condensed Nearest Neighbor + Tomek's modification of Condensed Nearest Neighbor (CNNTL). CNNTL applies the CNN method and the Tomek Links method in a chain to delete the instances that lead us to misclassify new instances in the imbalanced domains (Batista et al., 2004).
- *Neighborhood Cleaning Rule* (NCL). NCL finds a subset S of the training set T applying the neighborhood cleaning rule of examples (Laurikkala, 2001).
- *One Side Selection* (OSS). OSS finds a subset S of the training set T applying the OSS of examples (Kubat and Matwin, 1997).
- *Random Undersampling* (RUS). RUS deletes the majority of class data instances randomly (Batista et al., 2004).

- *Undersampling based on clustering* (SBC). After dividing all the training samples into some clusters, SBC selects a suitable number of majority class samples from each cluster by considering the ratio of the number of majority class samples to the number of minority class samples in the cluster (Yen and Lee, 2006).

For both oversampling and undersampling methods, the default parameters were used. The corresponding parameters set can be found in the method library of KEEL software (UGR Granada, Spain) (Alcalá-Fdez et al., 2011).

Classification

To classify and compare the different resampled datasets, 10 different machine learning algorithms, belonging to the family of supervised classification, and most used in the contest of neurophysiological signal processing, were applied as follows:

1. Decision tree (DT): coarse tree, whose maximum number of branch points is set to 4. The method adopts the Gini's diversity index as the split criterion and envisages a pruning procedure.
2. Discriminant analysis (DA): creates non-linear boundaries between the classes (quadratic discriminant analysis).
3. Logistic regression (LR).
4. Naïve Bayes (NB): the method supports continuous attributes by assuming a Gaussian distribution (Gaussian Naïve Bayes).
5. Support vector machine (SVM): characterized by coarse distinctions between the classes, with kernel scale set to $4\sqrt{P}$, where P is the number of predictors (Coarse Gaussian SVM).
6. KNN (K-nearest neighbors): where we set the number of neighbors to 100 (Coarse distinctions between classes) and used the Euclidean distance metric (coarse KNN).
7. Boosted Ensemble (EnsBO): ensemble classifier which uses the meta-algorithm AdaBoost (Freund and Schapire, 1999).
8. Bagged Ensemble (EnsBA), Random forest Bag, with DT learners. This implementation uses Breiman's "random forest" algorithm (Breiman, 2001).
9. Discriminant Analysis Ensemble (EnsDA): combines different feature subsets to improve the classification performance (subspace ensemble), and uses Discriminant learners.
10. KNN ensemble (EnsKNN): Subspace ensemble with Nearest Neighbor learners.

During the training phase, the validation step was performed through a 5-fold cross-validation approach. For all the considered methods, default parameters were used. The corresponding parameters set can be found in the *Matlab classification learner* toolbox specification.

Ensemble Methods for Imbalanced Domain

Since the main objective of the study was to compare the effect of different resampling techniques on the classifier performances, in the previous section we described both the standard and classical ensemble classifiers, with the resampling procedure applied before the classification.

However, in the past years, ensemble-based classifiers have been considered a suitable approach in the imbalanced domain, leading to the implementation of specific modification of the ensemble algorithm, in which the data rebalancing pre-processing is integrated into the ensemble algorithm and done before the learning stage of each classifier of the ensemble (Chawla et al., 2003; Seiffert et al., 2010). For this reason, we also tested six of these approaches, three belonging to boosting (methods 1–2–3) and three to bagging (methods 4–5–6) approach:

1. DATABoost: it combines the AdaBoost algorithm with a data generation strategy. It first identifies hard examples (seeds) and then carries out a rebalance process, always for both the classes (Guo and Viktor, 2004).
2. RUSBoost: multi-class AdaBoost with RUS in each iteration (Seiffert et al., 2010).
3. SMOTEBoost: multiclass AdaBoost with SMOTE in each operation (Chawla et al., 2003).
4. OVERBag: bagging with oversampling of the minority class (Wang and Yao, 2009).
5. SMOTEBag: bagging where SMOTE quantity of each bag varies (Wang and Yao, 2009).
6. UnderBag: bagging with undersampling of the majority class (Barandela et al., 2003b).

Performances Metrics

In common practice, accuracy is the most used measure to assess classifier performance. However, since it does not allow to distinguish between the number of correctly classified instances of the two different classes, it can lead to an erroneous conclusion when applied in the context of imbalanced datasets.

To assess and compare the performances of the classifiers, we used the following four metrics, which have been proven to be suitable for the imbalanced domain (Bekkar et al., 2013; López et al., 2013; Fernández et al., 2018):

$$AUC = \frac{1 + TPr + FPr}{2}$$

$$Fm = \frac{(1+\beta^2)(PPV \cdot TPr)}{\beta^2 \cdot PPV + TPr)}$$

$$GMean = \sqrt{\frac{TP}{TP+FN} \cdot \frac{TN}{FP+TN}}$$

$$BalACC = \frac{TPr + TNr}{2}$$

Where TPr is the *true positive rate* (or *sensitivity)*, TNr is the *true negative rate* (or *specificity)*, and PPV is the *positive predicted value*, respectively, defined as:

$$TPr = \frac{TP}{TP+FN}$$

$$TNr = \frac{TN}{TN+FP}$$

$$PPV = \frac{TP}{TP+FP}$$

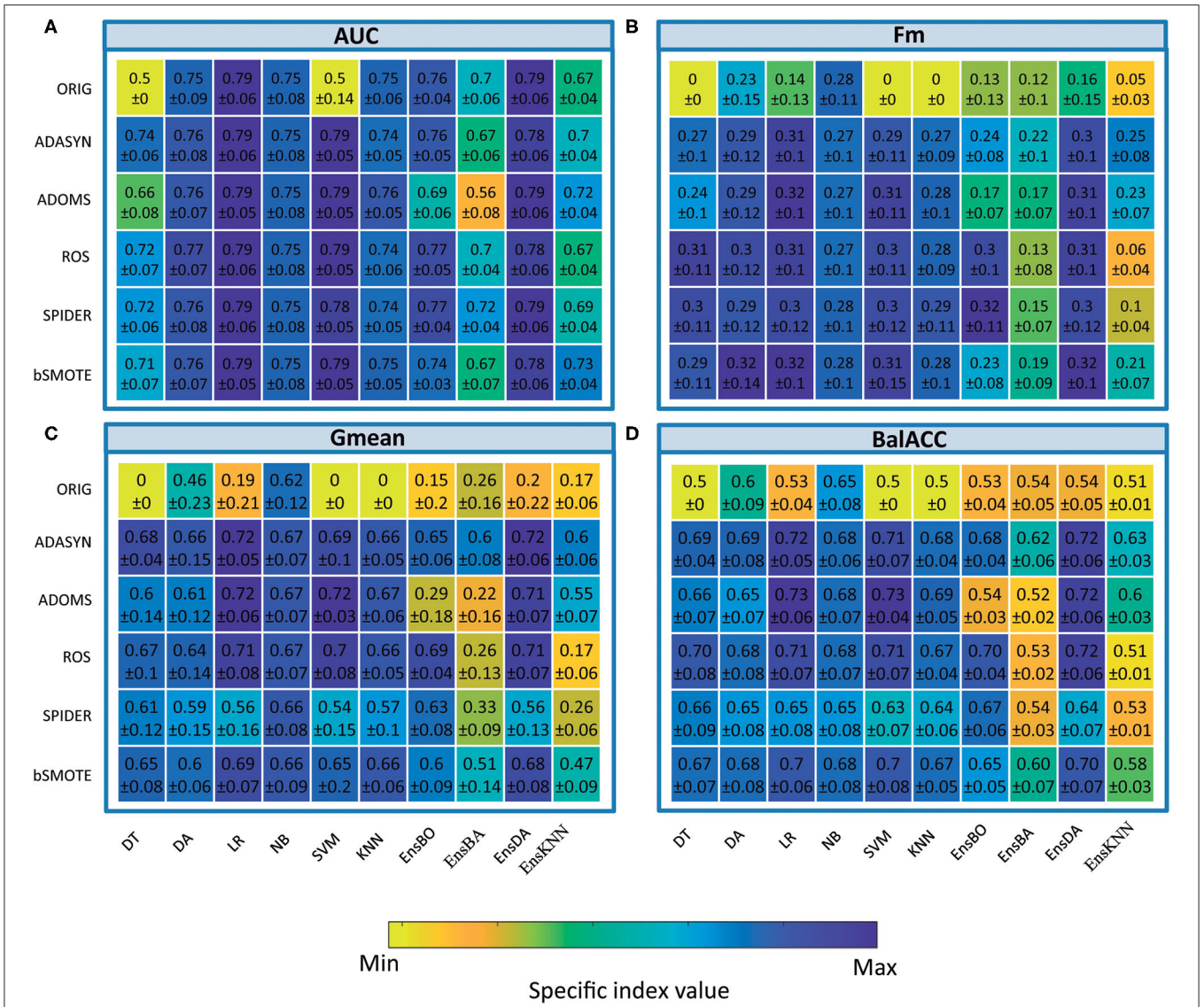


FIGURE 2 | Comparison of classifier performances (mean and SD values), among the different oversampling procedures, in terms of indices Area under the ROC curve (AUC) **(A)**, F-measure (Fm) **(B)** Geometric Mean (Gmean) **(C)**, and Balanced Accuracy (BalACC) **(D)**. *X*-axis: classifiers applied; *Y*-axis: resampling techniques. For ease of understanding, the colormap spans from minimum to maximum values of each specific index. As shown in **Table 2** and **Supplementary Tables 2–5** for statistical comparisons among these values.

Note that TP, TN, FP, and FN stay for true positives, true negatives, false positives, and false negatives, respectively. For Fm we used $\beta = 1$, to assign equal importance to both TP and PPV.

All the analyses were performed using the *KEEL* software (Alcalá-Fdez et al., 2011) and the *Matlab classification learner* toolbox.

Statistical Analysis

To compare the different resampling techniques, Friedman's test was applied to the four performances metrics AUC, Fm, Gmean, and BalACC (Friedman, 1937). When a significant difference among the group was found, Shaffer's *post-hoc* test was applied for multiple comparisons (Shaffer, 1986). The alpha level for statistical significance was set at 0.05, and the final adjusted *p*-values are used for the results. All the statistical comparisons were performed using SPSS (IBM Corp. Version 26.0. Armonk, NY, USA) and KEEL software.

Data are available from the corresponding authors upon request.

RESULTS

Oversampling

The average predicted performances in terms of AUC, Fm, Gmean, and BalACC are shown in **Figure 2**. For all 10 classifiers, the statistical results of the Friedman's Test and related

TABLE 3 | Friedman's and *post-hoc* Shaffer's test for the *oversampling* techniques applied to the four performance measures: Area under the ROC Curve (AUC), F-measure (Fm), Geometric Mean (Gmean), and Balanced Accuracy (BalACC).

Oversampling		*DT*	*DA*	*LR*	*NB*	*SVM*	*KNN*	*EnsBO*	*EnsBA*	*EnsDA*	*EnsKNN*
Original vs. *ADASYN*	AUC	–				–		–			–
	Fm	–						–	–		–
	Gmean	–	–	–		–	–	–	–	–	–
	BalACC	–	–	–	–	–	–	–	–	–	–
Original vs. *ADOMS*	AUC					–			+		–
	Fm			–		–	–				
	Gmean	–		–		–	–			–	–
	BalACC	–		–	–	–	–			–	–
Original vs. *ROS*	AUC	–				–					
	Fm	–		–		–	–	–		–	–
	Gmean	–	–	–		–	–	–		–	
	BalACC	–	–	–	–	–	–	–		–	
Original vs. *SPIDER*	AUC	–									
	Fm	–		–		–	–	–		–	–
	Gmean							–			
	BalACC				–			–			
Original vs. *bSMOTE*	AUC					–					–
	Fm	–		–		–	–			–	–
	Gmean	–	–	–		–	–	–	–	–	–
	BalACC	–	–	–		–	–	–		–	–
ADASYN vs. *ADOMS*	AUC							+			
	Fm										
	Gmean							+	+		
	BalACC							+	+		
ADASYN vs. *ROS*	AUC										
	Fm								+		
	Gmean								+		+
	BalACC								+		+
ADASYN vs. *SPIDER*	AUC										
	Fm										
	Gmean			+						+	+
	BalACC			+						+	+
ADASYN vs. *bSMOTE*	AUC										
	Fm										
	Gmean										
	BalACC										
ADOMS vs. *ROS*	AUC							–	–		+
	Fm							–			
	Gmean							–			+
	BalACC							–			+
ADOMS vs. *SPIDER*	AUC							–	–		
	Fm							–			
	Gmean			+			+			+	
	BalACC			+			+	–		+	
ADOMS vs. *bSMOTE*	AUC										
	Fm										
	Gmean								–		
	BalACC								–		

(Continued)

TABLE 3 | Continued

Oversampling		*DT*	*DA*	*LR*	*NB*	*SVM*	*KNN*	*EnsBO*	*EnsBA*	*EnsDA*	*EnsKNN*
ROS vs. *SPIDER*	AUC										
	Fm										
	Gmean			+						+	
	BalACC			+						+	
ROS vs. *bSMOTE*	AUC										–
	Fm										
	Gmean								–		–
	BalACC										–
SPIDER vs. *bSMOTE*	AUC										
	Fm										
	Gmean										
	BalACC										

The 10 columns refer to the 10 classifiers models. The comparisons showing significant results are indicated with a "–" sign when the first algorithm (of the two compared in each row) was lower or with a "+" sign when it was higher than the second one. The rows without significant differences are not reported. Complete results with the p-values can be found in ***Supplementary Tables 3–6****.*

Shaffer's *post-hoc* comparisons for AUC (a), Fm (b), Gmean (c), and BalACC (d) are shown in **Table 3**. Shaffer's *post-hoc* comparisons have been indicated only when Friedman's test resulted significantly. The sign "–" (respectively, "+") indicates that the first algorithm has a lower (higher) value than the second one.

- *The area under the ROC curve*: Friedman's test revealed significant differences among the pre-processing techniques only in five of the classifiers tested (DT, SVM, Ens_BO, Ens_BA, and Ens_KNN). For the two standard classifiers (DT and SVM), the *post-hoc* comparisons revealed differences only with respect to the original datasets, while no differences were present among the five oversampling techniques. Interestingly, for three of the four classical ensemble classifiers, none of the resampling techniques performed better than the original dataset. On the contrary, the ADOMS approach showed significantly lower AUC values than the other methods in both boosted and bagged ensemble classifiers. In the KNN ensemble, both original and ROS datasets reported the lowest performances (as shown in **Table 3** and **Supplementary Table 2**).
- *F-measure*: the significant differences have been revealed in 8 out of the 10 classifiers (DT, LR, SVM, KNN, EnsBO, EnsBA, EnsDA, and EnsKNN). The *post-hoc* comparisons showed the lower performance of the original dataset with respect to all resampling procedures in the six standard classifiers. In the ensemble both original and ADOMS had significantly lower Fm values than the other algorithm ms (as shown in **Table 3** and **Supplementary Table 3**).
- *Geometric Mean*: this metric exhibited more differences among the considered resampling approaches. All the classifiers except LR showed significant differences among the rebalancing approaches. In the standard classifiers and the EnsDA, the algorithm ADASYN, ADOMS, ROS, and bSMOTE performed better than both the original and SPIDER dataset. As for Fm, in Boosted and Bagged and KNN Ensemble ADOMS algorithm reported the lowest performance (as shown in **Table 3** and **Supplementary Table 4**).
- *Balanced Accuracy*: significant differences among the different resampling algorithms emerged for all the 10 classifiers. According to Shaffer's *post-hoc* analysis, ADASYN, ADOMS, ROS, and bSMOTE reported better performances than the original and SPIDER datasets in the standard classifiers. In the EnsBO and EnsBA, no differences were found between the original and ADOMS data set, which performed worse than the other resampling procedures. In the EnsDA classifier, the resampling algorithms ADASYN, ADOMS, ROS, and bSMOTE showed higher BalACC than the original and SPIDER dataset. In EnsKNN classifier, showed similar results than EnsDA, except for ROS, which reported BalACC comparable with original and SPIDER (as shown in **Table 3** and **Supplementary Table 5**).

Undersampling

The average predicted performances of undersampling procedures in terms of AUC, Fm, Gmean, and BalACC are shown in **Figure 3**. For all the 10 classifiers, the statistical results of the Friedman's Test and related Shaffer's *post-hoc* comparisons for AUC (a), Fm (b), Gmean, (c), and BalACC (d) are shown in **Table 4**, respectively. Shaffer's *post-hoc* comparisons have been indicated only when Friedman's Test resulted significantly; The sign "–" (respectively "+") indicates that the first algorithm has a lower (higher) value than the second one.

- The area under the ROC curve: significant differences among the pre-processing techniques are found in five of the classifiers tested (DT, SVM, KNN, EnsBA, and EnsKNN). In the DT classifier, all undersampling algorithms performed equally and better than the original one; in SVM, RUS,

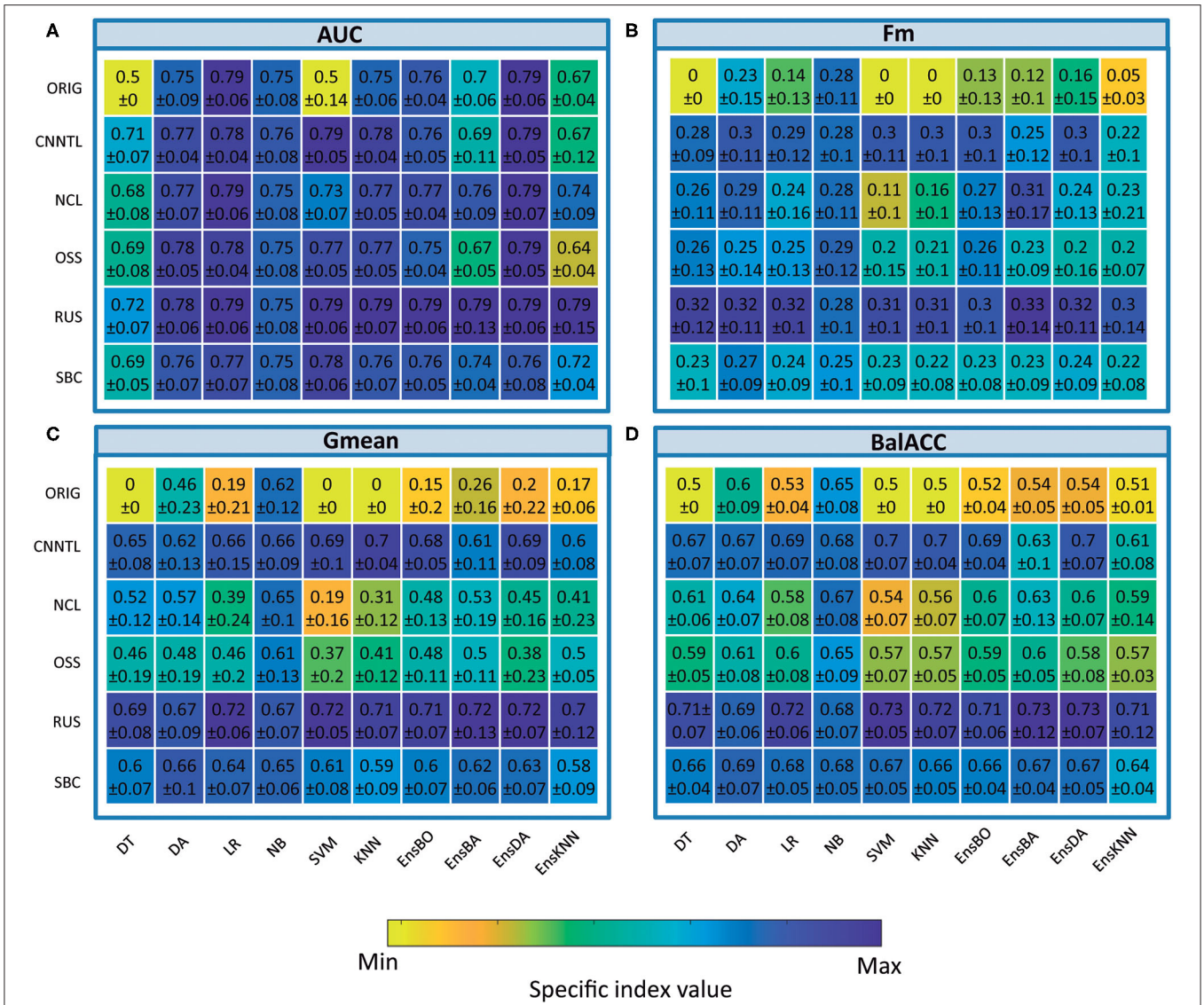


FIGURE 3 | Comparison of classifier performances (mean and SD), among the different undersampling procedures, in terms of indices AUC **(A)**, Fm **(B)**, Gmean **(C)**, and BalACC **(D)**. *X*-axis: classifiers applied; *Y*-axis: resampling techniques. For ease of understanding, the colormap spans from minimum to maximum values of each specific index. As shown in **Table 3** and **Supplementary Tables 6–9** for statistical comparisons among these values.

and CNNTL performed better than the others, and in KNN only RUS showed improved AUC performances with respect to the original and all the other resampling techniques. In EnsBA and EnsKNN, significantly improved performances were achieved by NCL, RUS, and SBC (as shown in **Table 4** and **Supplementary Table 6**).

- *F-measure*: Friedman's test revealed significant differences in 9 out of the 10 classifiers (all except NB). For standard classifiers, *post-hoc* comparisons showed the lower performance of the original dataset with respect to all resampling procedures except for SBC in DT classifier, NCL in SVM and KNN, and NCL, OSS, and SBC in LR classifier. As well as in standard classifiers, also in all the ensembles, the best performances were achieved by RUS, followed by the CNNTL algorithm (as shown in **Table 4** and **Supplementary Table 7**).
- *Geometric Mean* showed significant differences among the considered approaches for all the classifiers, proving to be more suited than AUC and Fm in capturing the differences among the resampling approaches. RUS, SBC, and CNNTL showed the highest performances, with significantly higher Gmean than the original dataset in all the classifiers except NB. Moreover, RUS indicated significantly higher performances than NCL and OSS (**Table 4** and **Supplementary Table 8**).
- *Balanced Accuracy* showed very similar patterns with respect to Gmean, denoting differences for all the classifiers. According to Shaffer's *post-hoc* analysis, CNNTL, RUS,

and SBC perform significantly better than the original dataset and the NCL and OSS resampling approaches, being RUS the best algorithms (as shown in **Table 4** and **Supplementary Table 9**).

Ensemble Methods for Imbalanced Domain

To compare the ensemble methods, we considered the two indices Gmean and BalACC since they have been shown to better capture the differences among the algorithms, as reported in the previous section.

Figure 4 shows the average ranking value for each of the proposed ensemble approaches, for both Gmean and BalACC. Corresponding results according to *post-hoc* Shaffer's test, comparing the seven approaches (original dataset and six ensembles) can be found in **Table 5**. According to Friedman's test, both the measures indicated significant differences among these techniques (Gmean: $p < 0.00001$; BalACC: $p < 0.00001$). A *post-hoc* analysis pointed out that DATABoost and SMOTEBag did not improve the performances with respect to the original dataset, and that SMOTEBoost, OVERBag showed higher BalACC than the original data but no differences in terms of Gmean. On the contrary, RUSBoost and UNDERBag showed significantly better performances than all the other algorithms, being UNDERBag the best one (**Table 5**).

Since in the previous section we used classical ensemble classifiers combined with a rebalancing pre-processing step, we also compared the one with better performances (EnsDA, after ADASYN and RUS resampling) with the best algorithm of the modified ensemble family UNDERBag. Interestingly, EnsDA, with both ADASYN and RUS pre-processing, showed significantly higher Gmean and BalACC than the UNDER_Ba approach ($p < 0.00519$ for ADASYN+Ens_DA vs. UNDERBag, and $p < 0.00104$ for RUS+Ens_DA vs. UNDERBag, for both Gmean and BalACC). **Figure 5** represents the comparison among these three methods, expressed in terms of ranking values.

Sensitivity and Specificity

To clarify the effective use of the proposed approach to EZ identification, we reported sensitivity and specificity for the different techniques tested in the study. Since ensemble approaches showed significantly lower performances than resampling in terms of performances metrics (as indicated in the previous paragraph), only the sensitivity and specificity of the latter were further analyzed. **Figure 6** shows the boxplots indicating the values of sensitivity (full-color boxes) and specificity (horizontal lines boxes) for the original dataset compared with the five oversampling (**Figure 6A**) and the five undersampling approaches (**Figure 6B**). Each box represents the variability among the 10 classification models. All sensitivity and specificity values are reported in **Table 6**. Such results confirmed the main evidence obtained by the other performance metrics: (i) original data were not able to provide a good classification, since all the models tended to classify the whole set of leads as non-EZ (sensitivity $\approx$ 0; specificity $\approx$ 1), confirming the biased classification toward the majority non-EZ class; (ii) oversampling improved classification performances, especially in terms of sensitivity. The Adasyn method provided the highest combination of both values (sensitivity and specificity $>$0.7) and the lowest variability of performances among the classification models. The ADOMS method showed average performances comparable with ADASYN, but much more variability with respect to the model choice. The SPIDER method was the least effective approach to improve the performances; (iii) Some undersampling approaches improved the classification performances, but with a strong variability among the different methods. NCL and OSS show results comparable to the original dataset. The RUS method provided the highest values of both sensitivity and specificity, comparable with the ADASYN approach. Interestingly, the SBC showed the highest sensibility values ($\approx$0.9), even if associated with a less balanced specificity.

Figure 7 shows the visualization of the surgical 3D scene for a representative patient (pt2), such as an indication of the resected zone (blue area), true EZ and non-EZ leads, and the EZ and non-EZ classification provided by the RUS + EnsDA method.

DISCUSSION

Machine learning approaches are being increasingly applied to the field of epilepsy, and specifically in the different datasets from neurophysiological recordings (Abbasi and Goldenholz, 2019). In this context, it is quite common to cope with the imbalanced datasets characterized by uneven distribution between majority and minority classes, which can lead to worse classification performances.

This is the case of the EZ localization in the pre-surgical planning to achieve seizure freedom after surgical resection of the EZ. One assessed clinical practice is the exploration through intracranial EEG recordings (SEEG) (Cardinale et al., 2019) combined with the visual analysis and advanced signal processing methods able to extract quantitative indexes to support the correct EZ localization (Bartolomei et al., 2017).

Intentionally, to sample a wide region of the epileptic brain, the explored brain regions are much wider than the true EZ, thus resulting in an imbalanced class distribution between EZ and non-EZ contacts, with the EZ being the most important class to be correctly identified to reduce or remove seizures, being the minority class. This led the classifier to be biased toward the majority (non-EZ) class.

Starting from the evidence that network analysis of interictal SEEG recordings could be very useful in support of the EZ localization (Varotto et al., 2012; Vlachos et al., 2017; Lagarde et al., 2018), in this study we demonstrated that the combination of supervised machine learning with appropriate data resampling approach can strongly improve its potential. For this reason, the idea of applying resampling techniques in the field of EZ localization should be taken into consideration.

At present, no study investigated the effect of imbalance domains on the performance of EZ localization methodologies. The previous studies demonstrated that the application of rebalancing techniques could strongly improve the classification of EEG signals for epilepsy diagnosis (Haldar et al., 2019; Kaur

TABLE 4 | Friedman's and *post-hoc* Shaffer's test for the *undersampling* techniques applied to the four performance measures: AUC, Fm, Gmean, and BalACC.

Undersampling		*DT*	*DA*	*LR*	*NB*	*SVM*	*KNN*	*EnsBO*	*EnsBA*	*EnsDA*	*EnsKNN*
Original vs. *CNNTL*	AUC	–				–					
	Fm	–		–		–	–	–		–	–
	Gmean	–	–	–		–	–	–	–	–	–
	BalACC	–	–	–	–	–	–	–		–	–
Original vs. *NCL*	AUC	–									
	Fm	–						–	–	–	
	Gmean										
	BalACC										
Original vs. *OSS*	AUC	–									
	Fm	–				–	–				–
	Gmean										
	BalACC										
Original vs. *RUS*	AUC	–				–	–				–
	Fm	–		–		–	–	–	–	–	–
	Gmean	–	–	–		–	–	–	–	–	–
	BalACC	–	–	–		–	–	–	–	–	–
Original vs. *SBC*	AUC	–				–					
	Fm					–	–				–
	Gmean	–		–		–	–	–	–	–	–
	BalACC	–		–		–	–	–	–	–	–
CNNTL vs. *NCL*	AUC					+			–		–
	Fm					+	+				–
	Gmean			+		+	+	+		+	
	BalACC			+		+	+				
CNNTL vs. *OSS*	AUC										
	Fm										
	Gmean				+	+	+	+		+	
	BalACC		+		+	+	+	+		+	
CNNTL vs. *RUS*	AUC								–		–
	Fm										
	Gmean										
	BalACC										
CNNTL vs. *SBC*	AUC								–		–
	Fm							+			
	Gmean										
	BalACC										
NCL vs. *OSS*	AUC								+		+
	Fm										
	Gmean										
	BalACC										
NCL vs. *RUS*	AUC					–					
	Fm					–	–				–
	Gmean	–		–		–	–	–	–	–	–
	BalACC	–		–		–	–	–	–	–	–
NCL vs. *SBC*	AUC										
	Fm										
	Gmean										
	BalACC			–							

(Continued)

TABLE 4 | Continued

Undersampling		*DT*	*DA*	*LR*	*NB*	*SVM*	*KNN*	*EnsBO*	*EnsBA*	*EnsDA*	*EnsKNN*
OSS vs. *RUS*	AUC								–		–
	Fm										
	Gmean	–	–	–		–	–	–	–	–	–
	BalACC	–	–	–		–	–	–	–	–	–
OSS vs. *SBC*	AUC								–		–
	Fm										
	Gmean										
	BalACC										
RUS vs. *SBC*	AUC										
	Fm			+							
	Gmean										
	BalACC										

The 10 columns refer to the 10 classifiers models. The comparisons showing significant results are indicated with a "–" sign when the first algorithm (of the two compared in each row) was lower or with a "+" sign when it was higher than the second one. The rows without significant differences are not reported. Complete results with the p-values can be found in ***Supplementary Tables 3–6****.*

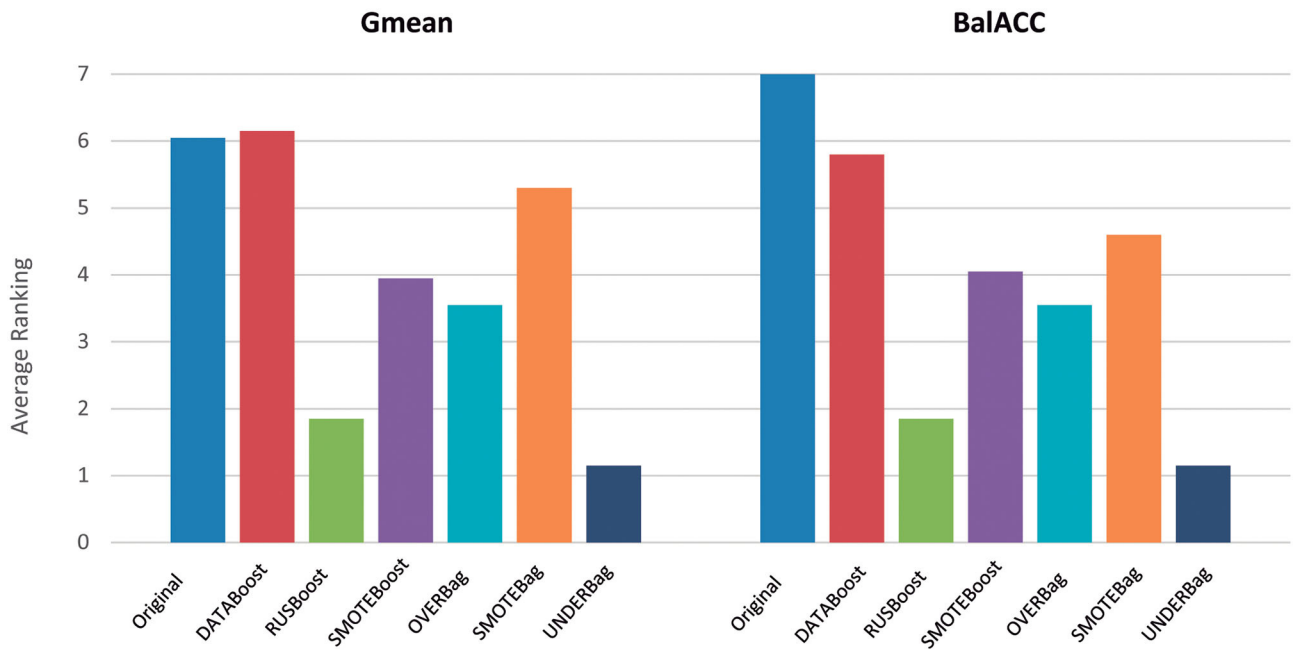


FIGURE 4 | Comparison of performances among the original dataset (blue bars) and six modified ensemble approaches for the imbalanced domain, in terms of the ranking (*y*-axis) of Gmean and BalACC. Lower-ranking values indicate better performances. As shown in **Table 4**, for the results of statistical comparisons.

et al., 2020) and automatic seizure detection (Cosgun et al., 2019; Romaissa et al., 2019; Masum et al., 2020). However, in most of them, the well-known and assessed resampling techniques belonging to the SMOTE family were applied, and systematic comparison with other possible approaches was missing.

In this study, we compared five oversampling and five undersampling procedures and tested the resulting rebalanced datasets with 10 different machine learning classifiers. Moreover, we also tested six specific ensemble methods properly modified for imbalanced domain and belonging to data variation-based ensemble.

Our study focuses on identifying the best resampling and classification approach to support the classification of brain regions as EZ or non-EZ, using the indexes derived from connectivity and graph-theory analysis of interictal SEEG recording as features. The selection of the nine graph-theory-based indexes used as input features of the classifiers was based on the preliminary analysis we performed, showing that the combination of these indexes was the most appropriate

TABLE 5 | Shaffer's test for the ensemble approaches for the imbalance domain.

Ensemble	AUC		Fm		Gmean		BalACC	
Original vs. *DATABoost*	–	0.000	–	0.014		1.895		1.285
Original vs. *RUSBoost*		2.344		0.789	–	0.000	–	0.000
Original vs. *SMOTEBoost*		0.165		0.423		0.297	–	0.034
Original vs. *OVERBag*		0.555		1.285		0.106	–	0.005
Original vs. *SMOTEBag*	–	0.040	–	0.006		1.895		0.143
Original vs. *UNDERBag*		2.344	–	0.000	–	0.000	–	0.000
DATABoost vs. *RUSBoost*	+	0.000		0.789	–	0.000	–	0.001
DATABoost vs. *SMOTEBoost*		0.555		1.285		0.251		0.491
DATABoost vs. *OVERBag*		0.218		0.372		0.078		0.199
DATABoost vs. *SMOTEBag*		1.499		1.814		1.895		1.285
DATABoost vs. *UNDERBag*	+	0.000		0.298	–	0.000	–	0.000
RUSBoost vs. *SMOTEBoost*		0.165		1.814		0.297		0.205
RUSBoost vs. *OVERBag*		0.555		1.814		0.549		0.549
RUSBoost vs. *SMOTEBag*	+	0.040		0.701	+	0.005	+	0.049
RUSBoost vs. *UNDERBag*		2.344	–	0.002		1.895		1.406
SMOTEBoost vs. *OVERBag*		2.344		1.516		1.895		1.406
SMOTEBoost vs. *SMOTEBag*		2.344		1.031		0.974		1.406
SMOTEBoost vs. *UNDERBag*	–	0.024	–	0.006		0.056	–	0.034
OVERBag vs. *SMOTEBag*		1.663		0.298		0.491		1.285
OVERBag vs. *UNDERBag*		0.091	–	0.000		0.143		0.143
SMOTEBag vs. *UNDERBag*	–	0.003		0.423	–	0.000	–	0.005

Red color indicates the p-values with significant differences according to Shaffer's post-hoc ($p < 0.05$); the sign "–" (respectively "+") indicates that the first algorithm has a lower (higher) value than the second one.

to achieve the best EZ classification. In the contest of EZ localization, despite the early application of several other signal processing approaches for feature extraction, such as working in the frequency domain or by non-linear analysis, network analysis started only recently to be employed based on the evidence that focal epilepsy is a network disease. However, most of these recent network studies normally focus only on the connectivity analysis that is rarely combined with the pre-processing approaches, due to the huge amount of data to be processed. For this reason, in this study, we mainly focused on presenting pre-processing, in combination with a few of such feature extraction and connectivity measures in the literature, to provide evidence of and support for a proper pre-processing method in this context.

Regarding oversampling, all five approaches reported improved performances with respect to the original dataset. The differences among the five oversampling approaches varied according to the considered classifiers.

Adaptive Synthetic Sampling resulted to be the most robust approach among the classifiers. ADOMS was the less robust and most sensitive to the choice of classifier, being comparable or even slightly better than ADASYN in LR, SVM, KNN, EnsDA, and EnsKNN, while as bad as the original dataset in DA, EnsBO, and EnsBA. SPIDER was the least effective, with performances significantly worse than the other approaches and comparable with the original dataset for some classifiers, especially the classical ensemble family.

Regarding undersampling, all the approaches appeared to be less influenced by the classifier choice than the oversampling. Two of the proposed methods, NCL and OSS, did not improve the classification performances with respect to the original data. The other approaches were significantly better than original data, with RUS, the simplest of the proposed methods, being the best one.

Interestingly RUS showed higher, even not significant, performances than the best oversampling approach, ADASYN.

The resampling technique is not the only family to cope with the imbalanced domain. A wide number of approaches exist to deal with this problem, which can be mainly categorized as data-level or algorithmic-level approaches (López et al., 2013). Rebalancing belongs to the data-level approaches, in which data are pre-processed before the classification (Lee, 2014). On the contrary, in the algorithmic-level ones, the classification algorithm is modified to deal with the imbalanced nature (Barandela et al., 2003a). The cost-sensitive approaches combine both the data and algorithmic levels, by assigning different misclassification costs for the two classes and modify the classification algorithm to minimize the higher misclassification cost (Domingos, 1999; Zhou and Liu, 2006; Sun et al., 2007).

The main limitation of cost-sensitive approaches is the need of defining the correct misclassification costs for the two classes, which may not be so clear in many clinical problems, as in our case.

In this paper, we focused on the rebalancing techniques since they can be quite easily implemented, and are independent of the underlying classifiers, which can be an advantage in problems where the selection of the most appropriate classifier is not clear (Batista et al., 2004; Batuwita and Palade, 2010).

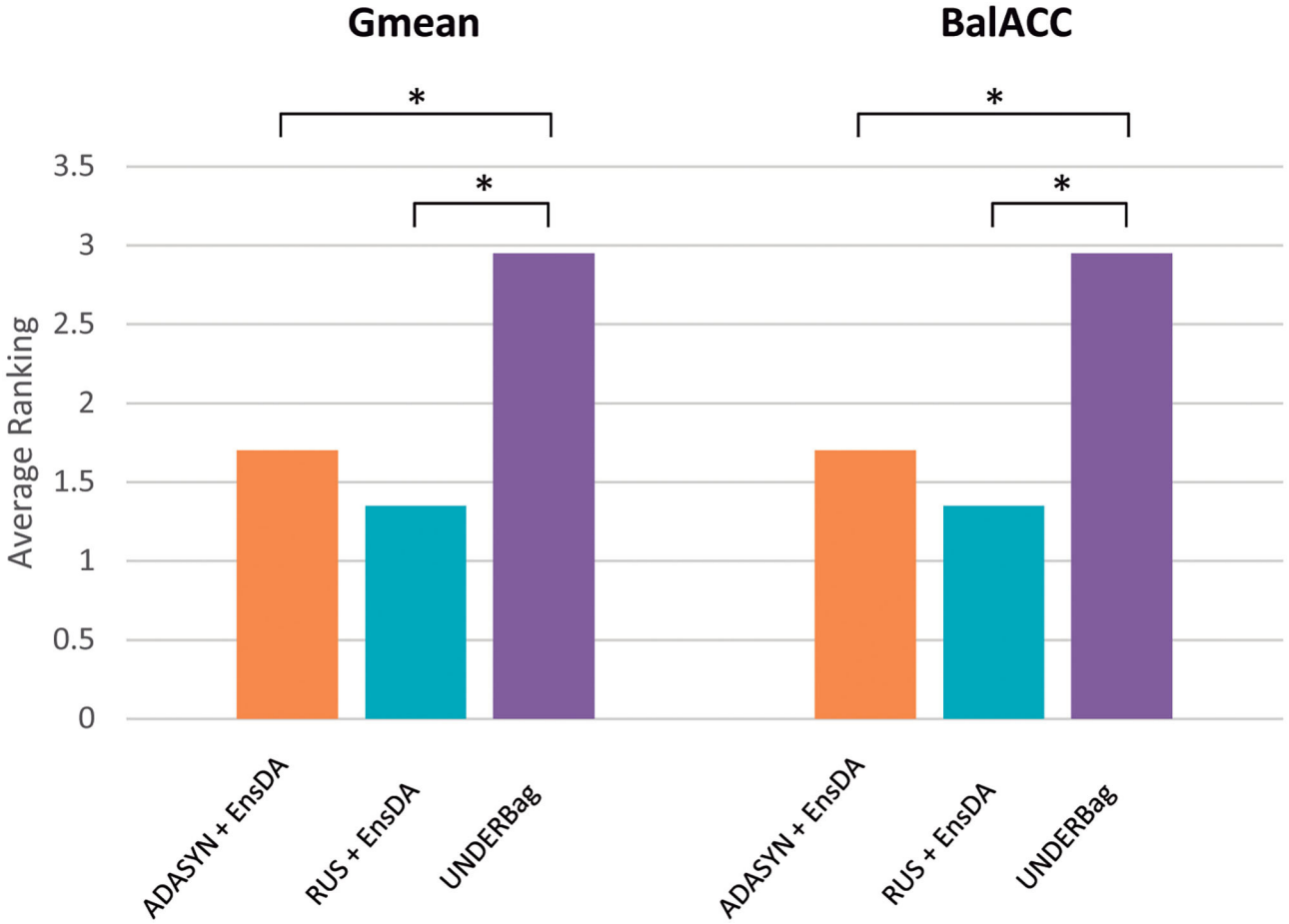


FIGURE 5 | A final comparison of performances among the best approach for each of the resampling families considered: (1) Oversampling: ADASYN + EnsDA [Adaptive Synthetic Sampling (ADASYN) combined with "classical" ensemble approach EnsembleDA; orange bars]; (2) Undersampling: RUS + EnsDA [Random Undersampling (RUS) combined with "classical" ensemble approach *EnsembleDA*—blue bars]; (3) Specific ensemble learner for imbalanced domain: UNDERBagging (violet bars). *Y*-axis represents ranking values, for both the Gmean and BalACC. The lower-ranking values indicate better performances. *indicates significant differences according to Shaffer's *post-hoc* analysis.

In addition, several modifications of ensemble methods for the imbalanced domain have been proposed (Rokach, 2010), both working at data-level approach, through the data pre-processing before each step of the ensemble classification (Breiman, 1996; Freund and Schapire, 1997; Kuncheva, 2014), or with algorithmic-level cost-sensitive modification (Sun et al., 2007).

As part of the data-level approaches, we considered and tested, in this study, six different data-level ensemble algorithms. As reported in a previous study (Galar et al., 2012), we found that the simplest algorithms, UNDERBag and RUSBoost emerged as the best ensemble methods, while offering lower computation costs.

Interestingly, when compared these results with those obtained by a standard single-step resampling approach combined with a classical ensemble algorithm, we found significantly higher performances in the latter family, in particular for the combination (ADASYN + EnsDA and RUS + EnsDA).

This highlights again that the simplest algorithms guarantee high performances, and that their very low computational complexity can be a strong advantage toward routine clinical applications.

It is important to notice that the performances of the different resampling techniques are strongly influenced by the choice of the classifier. This highlights that the selection of the resampling approach for a specific dataset should always take into consideration the choice of the classifier.

Regarding the measure to assess and compare the performances, in this study we applied four measures considered most appropriate to deal with imbalanced classification: AUC, Fm, Gmean, and BalACC (Bekkar et al., 2013). Several studies already highlighted that the choice of the proper evaluation measures for model assessment is one of the most complex issues faced in the imbalanced data learning context and how the application of more standard measures, such as accuracy, could lead to erroneous interpretations and biased classification (Weiss, 2004).

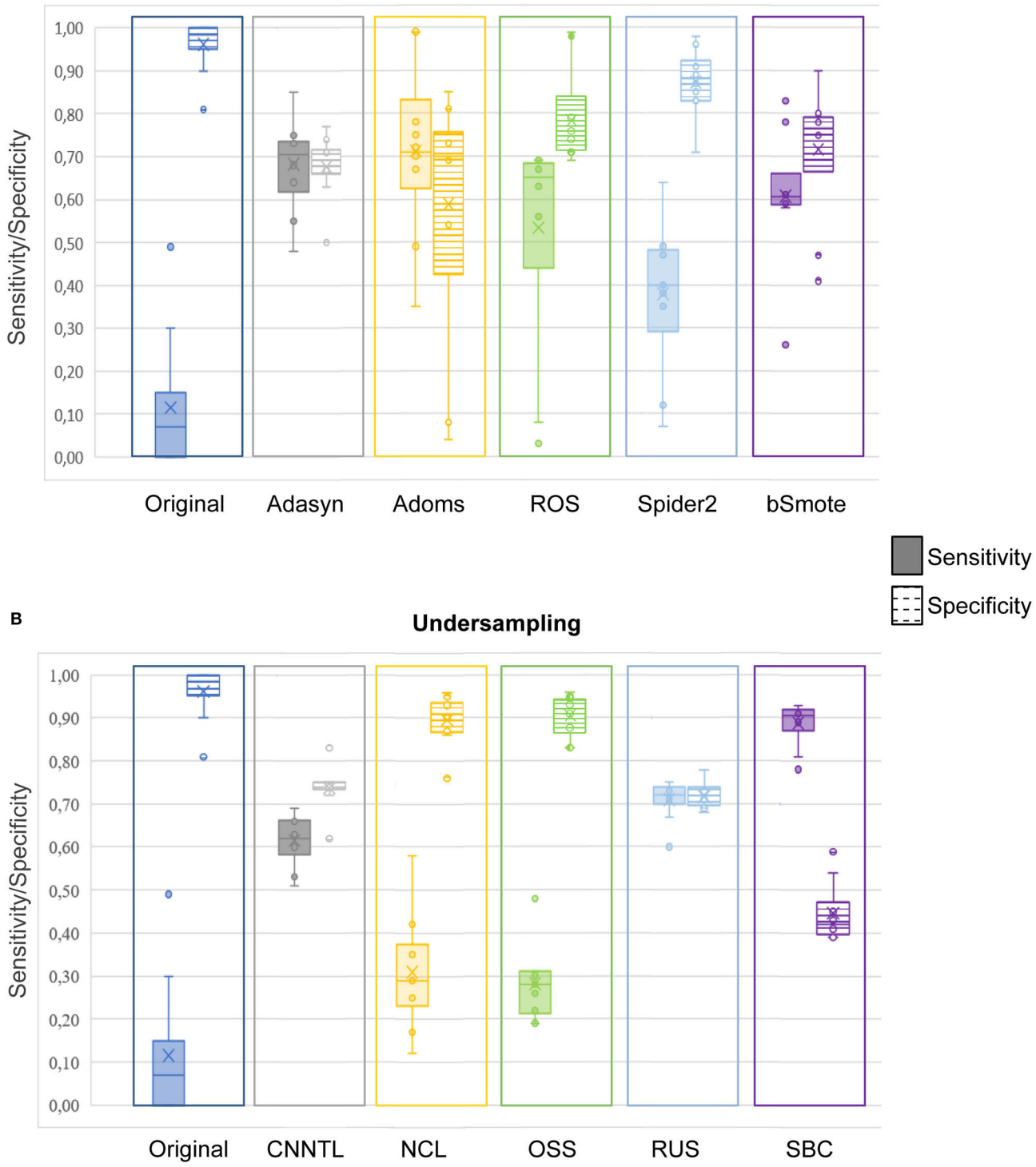


FIGURE 6 | Boxplot representing the values of sensitivity (full-color bars) and specificity (horizontal lines bars) for the original dataset and the five oversampling **(A)** and undersampling approaches **(B)**. Each box represents the variability among the 10 classification models. The middle line indicates the median value; upper and lower limits of the box indicate the first and third quartile; external points indicate outliers; x indicates the mean values. The corresponding values can be found in **Table 6**.

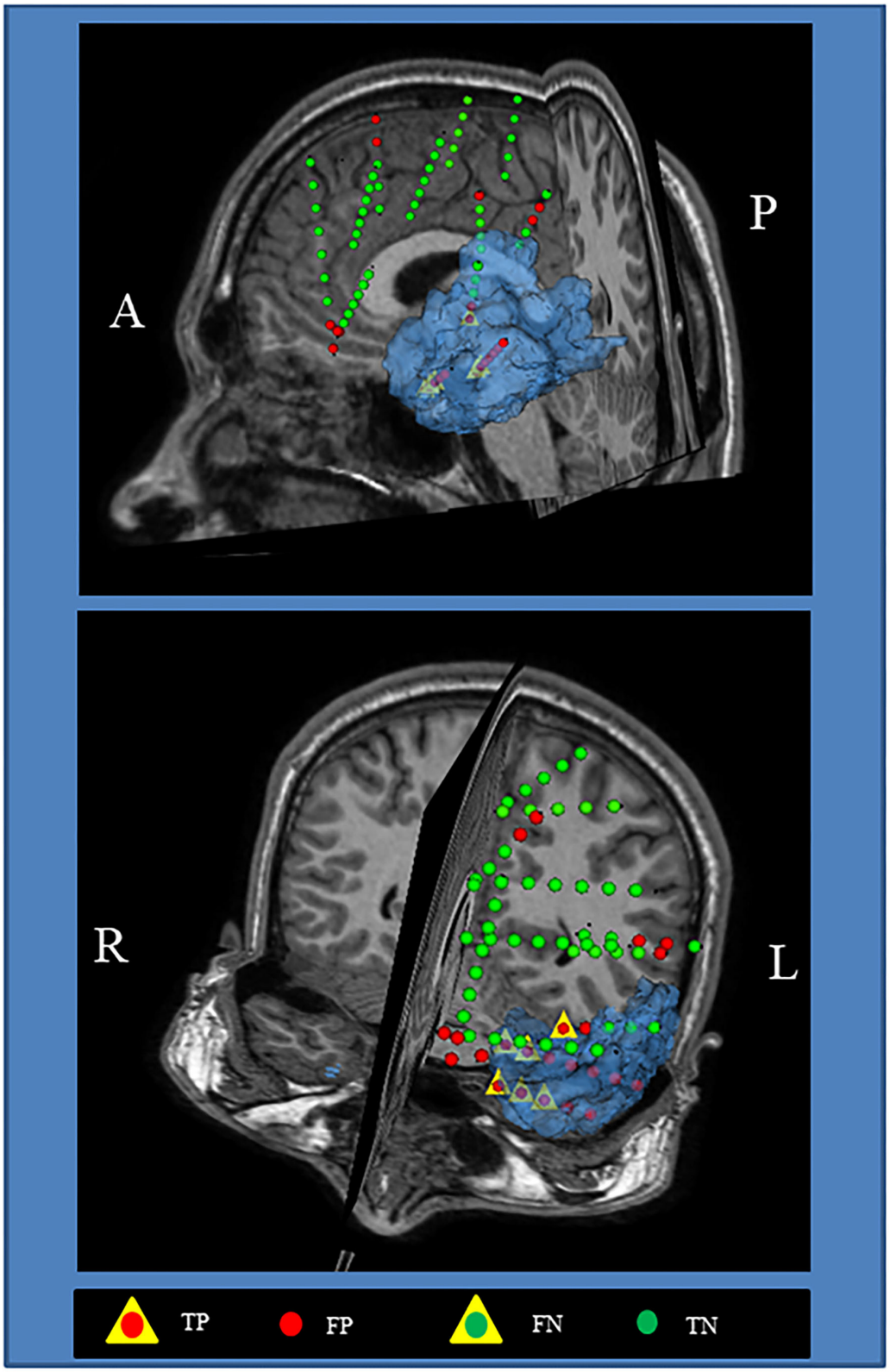


FIGURE 7 | Visualization of the surgical 3D scene for a representative patient (pt2). The Blue area indicates the final resected zone. Red and green dot points indicate leads classified as epileptogenic zone (EZ) A, anterior; L, left; P, posterior; R, right.

TABLE 6 | The Sensitivity (Sens) and Specificity (Spec) values for oversampling and undersampling techniques.

	Oversampling											
	Orig		**Adasyn**		**Adoms**		**ROS**		**Spider2**		**bSmote**	
	Sens	***Spec***	***Sens***	***Spec***	***Sens***	***Spec***	***Sens***	***Spec***	***Sens***	***Spec***	***Sens***	***Spec***
DT	0.00	1.00	0.75	0.63	0.78	0.54	0.63	0.76	0.49	0.83	0.59	0.75
DA	0.30	0.90	0.64	0.74	0.49	0.81	0.56	0.79	0.47	0.83	0.58	0.78
LR	0.08	0.98	0.73	0.71	0.72	0.74	0.68	0.74	0.40	0.89	0.61	0.79
NB	0.49	0.81	0.68	0.68	0.67	0.69	0.67	0.69	0.64	0.71	0.62	0.73
SVM	0.00	1.00	0.73	0.68	0.75	0.70	0.69	0.72	0.35	0.91	0.60	0.80
KNN	0.00	1.00	0.68	0.67	0.68	0.70	0.63	0.71	0.40	0.87	0.59	0.75
EnsBO	0.06	0.99	0.85	0.50	0.99	0.08	0.67	0.72	0.48	0.85	0.83	0.47
EnsBA	0.10	0.97	0.55	0.68	0.99	0.04	0.08	0.98	0.12	0.96	0.78	0.41
EnsDA	0.09	0.98	0.73	0.71	0.70	0.73	0.69	0.74	0.38	0.90	0.61	0.78
EnsKNN	0.03	0.99	0.48	0.77	0.35	0.85	0.03	0.99	0.07	0.98	0.26	0.90
Mean	**0.12**	**0.96**	**0.68**	**0.68**	**0.71**	**0.59**	**0.53**	**0.78**	**0.38**	**0.87**	**0.61**	**0.72**
St. Dev	***0.16***	***0.06***	***0.11***	***0.07***	***0.20***	***0.29***	***0.26***	***0.11***	***0.17***	***0.08***	***0.15***	***0.15***
	Undersampling											
	Orig		**CNNTL**		**NCL**		**OSS**		**RUS**		**SBC**	
	Sens	***Spec***	***Sens***	***Spec***	***Sens***	***Spec***	***Sens***	***Spec***	***Sens***	***Spec***	***Sens***	***Spec***
DT	0.00	1.00	0.61	0.74	0.35	0.87	0.28	0.91	0.71	0.71	0.92	0.39
DA	0.30	0.90	0.51	0.83	0.42	0.86	0.30	0.91	0.60	0.78	0.78	0.59
LR	0.08	0.98	0.63	0.75	0.26	0.91	0.28	0.92	0.71	0.74	0.91	0.45
NB	0.49	0.81	0.61	0.74	0.58	0.76	0.48	0.83	0.67	0.69	0.81	0.54
SVM	0.00	1.00	0.69	0.72	0.12	0.96	0.19	0.96	0.75	0.70	0.92	0.41
KNN	0.00	1.00	0.67	0.74	0.17	0.95	0.19	0.95	0.74	0.70	0.93	0.39
EnsBO	0.06	0.99	0.64	0.75	0.29	0.91	0.26	0.93	0.73	0.70	0.92	0.41
EnsBA	0.10	0.97	0.53	0.74	0.36	0.90	0.31	0.88	0.74	0.73	0.90	0.43
EnsDA	0.09	0.98	0.66	0.74	0.29	0.91	0.22	0.94	0.71	0.74	0.90	0.45
EnsKNN	0.03	0.99	0.60	0.62	0.25	0.93	0.31	0.83	0.74	0.68	0.89	0.40
Mean	**0.12**	**0.96**	**0.62**	**0.74**	**0.31**	**0.90**	**0.28**	**0.91**	**0.71**	**0.72**	**0.89**	**0.45**
St. Dev	***0.16***	***0.06***	***0.06***	***0.05***	***0.13***	***0.06***	***0.08***	***0.05***	***0.05***	***0.03***	***0.05***	***0.07***

Single values for each of the 10 classifier models, as well as mean and standard deviation (St.Dev.) are indicated.

These four measures provided complementary results and to properly evaluate the performances of different approaches, it is important to take into account the combination of them, especially considering which aspect is more important in the specific problem we are facing. Particularly, in this case, we noticed that AUC and Fm did not completely capture differences in the model performances. On the other side, as already described in another paper (Luque et al., 2019), Gmean and BalACC appear to be good performance metrics when the main focus is to maximize sensitivity, without losing too much specificity.

DATA AVAILABILITY STATEMENT

Data are available from the corresponding authors upon request. Requests to access these datasets should be directed to giulia.varotto@istituto-besta.it.

ETHICS STATEMENT

The study was approved by the Ethics Committee of the Fondazione IRCCS Istituto Neurologico Carlo Besta of Milan and was carried out in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All of the subjects gave their written informed consent before being included in the study.

AUTHOR CONTRIBUTIONS

GV: designed and conceptualized the study, analyzed and interpreted the data, and drafted the manuscript for intellectual content. GS: contributed to design the study, analyzed the data, and contributed to draft and revise the manuscript. LT and FG: major role in the acquisition of data and contributed to revise the manuscript. SF and FP: interpreted the data and contributed to

draft and revise the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the DESIRE (Strategies for Innovative Research to Improve Diagnosis, Prevention, and Treatment in children with difficult to treat epilepsy), an FP7 funded project (Grant Agreement No. 602531), from the European Commission, and the Grants Nos. RF-2011-02350578 and RF-2010-2319316 from the Italian Ministry of Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2021.715421/full#supplementary-material

Supplementary Figure 1 | A subset of stereo-electroencephalography (SEEG) traces recorded from pt2, and corresponding adjacency matrices for the first 3 of the 36 epochs analyzed.

Supplementary Table 1 | Description of the set of graph-theory based centrality measures used in this study.

Supplementary Table 2 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *AUC* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (*p*-values in the first line). Red color indicates *p*-values with significant differences according to shaffer *post-hoc* ($p < 0.05$); "–" (respectively "+") indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 3 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *Fm* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (*p*-values in the first line). Red color indicates *p*-values with significant differences according to shaffer *post-hoc* ($p < 0.05$); "–" (respectively "+") indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 4 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *Gmean* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (*p*-values in the first line). Red color indicates *p*-values with significant differences according to shaffer *post-hoc* ($p < 0.05$); "–" (respectively "+") indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 5 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *BalACC* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (*p*-values in the first line). Red color indicates *p*-values with significant differences according to shaffer *post-hoc* ($p < 0.05$); "–" (respectively "+") indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 6 | Friedman and *post-hoc* Shaffer test for the *oversampling* techniques with *AUC* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (*p*-values in the first line). Red color indicates *p*-values with significant differences according to shaffer *post-hoc* ($p < 0.05$); "–" (respectively "+") indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 7 | Friedman and *post-hoc* Shaffer test for the *undersampling* techniques with *Fm* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (*p*-values in the first line). Red color indicates *p*-values with significant differences according to shaffer *post-hoc* ($p < 0.05$); "–" (respectively "+") indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 8 | Friedman and *post-hoc* Shaffer test for the *underampling* techniques with *Gmean* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (*p*-values in the first line). Red color indicates *p*-values with significant differences according to shaffer *post-hoc* ($p < 0.05$); "–" (respectively "+") indicates that the first algorithm has lower (higher) value than the second one.

Supplementary Table 9 | Friedman and *post-hoc* Shaffer test for the *undersampling* techniques with *BalACC* measure. Shaffer *post-hoc* comparisons have been indicated only when Friedman test resulted significant (*p*-values in the first line). Red color indicates *p*-values with significant differences according to shaffer *post-hoc* ($p < 0.05$); "–" (respectively "+") indicates that the first algorithm has lower (higher) value than the second one.

REFERENCES

Abbasi, B., and Goldenholz, D. M. (2019). Machine learning applications in epilepsy. *Epilepsia* 60, 2037–2047. doi: 10.1111/epi.16333

Acharya, U. R., Hagiwara, Y., and Adeli, H. (2018). Automated seizure prediction. *Epilepsy Behav*. 88, 251–261 doi: 10.1016/j.yebeh.2018.09.030

Adkinson, J. A., Karumuri, B., Hutson, T. N., Liu, R., Alamoudi, O., Vlachos, I., et al. (2019). Connectivity and centrality characteristics of the epileptogenic focus using directed network analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 22–30. doi: 10.1109/TNSRE.2018.2886211

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., et al. (2011). KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *J. Mult. Log. Soft Comput.* 17, 255–287. Available online at: http://www.oldcitypublishing.com/journals/mvlsc-home/mvlsc-issue-contents/mvlsc-volume-17-number-2-3-2011/mvlsc-17-2-3-p-255-287/

Ali, A., Shamsuddin, S. M., and Ralescu, A. L. (2015). Classification with class imbalance problem: a review. *Int. J. Adv. Soft Comput. Appl.* 7, 176–204.

Armañanzas, R., Alonso-Nanclares, L., DeFelipe-Oroquieta, J., Kastanauskaite, A., de Sola, R. G., DeFelipe, J., et al. (2013). Machine learning approach for the outcome prediction of temporal lobe epilepsy surgery. *PLoS ONE*. 8:e62819. doi: 10.1371/journal.pone.0062819

Azami, M., El Hammers, A., Jung, J., Costes, N., Bouet, R., and Lartizien, C. (2016). Detection of lesions underlying intractable epilepsy on t1-weighted mri as an outlier detection problem. *PLoS ONE* 1:e0161498. doi: 10.1371/journal.pone.0161498

Barandela, R., Sánchez, J. S., García, V., and Rangel, E. (2003a). Strategies for learning in class imbalance problems. *Pattern Recognit.* 36, 849–851. doi: 10.1016/S0031-3203(02)00257-1

Barandela, R., Sánchez, J. S., and Valdovinos, R. M. (2003b). New Applications of ensembles of classifiers. *Pattern Anal. Appl.* 6, 245–256. doi: 10.1007/s10044-003-0192-z

Bartolomei, F., Lagarde, S., Wendling, F., McGonigal, A., Jirsa, V., Guye, M., et al. (2017). Defining epileptogenic networks: contribution of SEEG and signal analysis. *Epilepsia* 58, 1131–1147. doi: 10.1111/epi.13791

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* 6, 20–29. doi: 10.1145/1007730.1007735

Batuwita, R., and Palade, V. (2010). "Efficient resampling methods for training support vector machines with imbalanced datasets," in *Proceedings of the International Joint Conference on Neural Networks*, 1–8. doi: 10.1109/IJCNN.2010.5596787

Bekkar, M., Djemaa, H. K., and Alitouche, T. A. (2013). Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* 3, 2224–5782. Available online at: https://www.iiste.org/Journals/index.php/JIEA/article/view/7633

Breiman, L. (1996). Bagging predictors. *Mach. Learn.* 24, 123–140. doi: 10.1007/BF00058655

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Bulacio, J. C., Jehi, L., Wong, C., Gonzalez-Martinez, J., Kotagal, P., Nair, D., et al. (2012). Long-term seizure outcome after resective surgery in

patients evaluated with intracranial electrodes. *Epilepsia* 53, 1722–1730. doi: 10.1111/j.1528-1167.2012.03633.x

Cardinale, F., Rizzi, M., Vignati, E., Cossu, M., Castana, L., d'Orio, P., et al. (2019). Stereoelectroencephalography: retrospective analysis of 742 procedures in a single centre. *Brain* 142, 2688–2704. doi: 10.1093/brain/awz196

Chawla, N. V. (2009). "Data mining for imbalanced datasets: an overview," in *Data Mining and Knowledge Discovery Handbook*, eds O. Maimon, and L. Rokach (Boston, MA: Springer). doi: 10.1007/978-0-387-09823-4_45

Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. (2003). "SMOTEBoost: improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases. Lecture Notes in Computer Science, Vol. 2838*, eds N. Lavrač N, D. Gamberger, L. Todorovski, H. Blockeel (Berlin: Springer). doi: 10.1007/978-3-540-39804-2_12

Cosgun, E., Celebi, A., and Gullu, M. K. (2019). "Epileptic seizure prediction for imbalanced datasets," in *Medical Technologies Congress (TIPTEKNO)* (Izmir), 1–4. doi: 10.1109/TIPTEKNO.2019.8895137

Daoud, H., and Bayoumi, M. A. (2019). Efficient epileptic seizure prediction based on deep learning. *IEEE Trans. Biomed. Circuits Syst*. 13, 804–813. doi: 10.1109/TBCAS.2019.2929053

Dian, J. A., Colic, S., Chinvarun, Y., Carlen, P. L., and Bardakjian, B. L. (2015). Identification of brain regions of interest for epilepsy surgery planning using support vector machines. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc*. 2015, 6590–6593. doi: 10.1109/EMBC.2015.7319903

Domingos, P. (1999). "MetaCost: a general method for making classifiers cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data* (San Diego, CA), 155–164. doi: 10.1145/312129.312220

Elahian, B., Yeasin, M., Mudigoudar, B., Wheless, J. W., and Babajani-Feremi, A. (2017). Identifying seizure onset zone from electrocorticographic recordings: a machine learning approach based on phase locking value. *Seizure* 51, 35–42. doi: 10.1016/j.seizure.2017.07.010

Engel, J. (1993). Update on surgical treatment of the epilepsies: summary of the second international palm desert conference on the surgical treatment of the epilepsies (1992). *Neurology* 43, 1612–1617. doi: 10.1212/WNL.43.8.1612

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). "Foundations on imbalanced classification," in *Learning From Imbalanced Data Sets* (Cham: Springer). doi: 10.1007/978-3-319-98074-4_2

Fiest, K. M., Sauro, K. M., Wiebe, S., Patten, S. B., Kwon, C. S., Dykeman, J., et al. (2017). Prevalence and incidence of epilepsy. *Neurology* 88, 296–303. doi: 10.1212/WNL.0000000000003509

Frank, B., Hurley, L., Scott, T. M., Olsen, P., Dugan, P., and Barr, W. B. (2018). Machine learning as a new paradigm for characterizing localization and lateralization of neuropsychological test data in temporal lobe epilepsy. *Epilepsy Behav*. 86, 58–65. doi: 10.1016/j.yebeh.2018.07.006

Freund, Y., and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci*. 55, 119–139. doi: 10.1006/jcss.1997.1504

Freund, Y., and Schapire, R. E. (1999). A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.* 14, 771–780.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc*. 32:675. doi: 10.1080/01621459.1937.10503522

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," in *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), Vol. 42* (Institute of Electrical and Electronics Engineers Inc.), 463–484. doi: 10.1109/TSMCC.2011.21 61285

Gleichgerrcht, E., Munsell, B., Bhatia, S., Vandergrift, W. A., Rorden, C., McDonald, C., et al. (2018). Deep learning applied to whole-brain connectome to determine seizure control after epilepsy surgery. *Epilepsia* 59, 1643–1654. doi: 10.1111/epi.14528

Goldenholz, D. M., Jow, A., Khan, O. I., Bagić, A., Sato, S., Auh, S., et al. (2016). Preoperative prediction of temporal lobe epilepsy surgery outcome. *Epilepsy Res.* 127, 331–338. doi: 10.1016/j.eplepsyres.2016.09.015

Guo, H., and Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation. *ACM SIGKDD Explor. Newsl.* 6, 30–39 doi: 10.1145/1007730.1007736

Haldar, S., Mukherjee, R., Chakraborty, P., Banerjee, S., Chaudhury, S., and Chatterjee, S. (2019). "Improved epilepsy detection method by addressing class imbalance problem," in *IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON),* 934–939. doi: 10.1109/IEMCON.2018.8614826

Han, H., Wang, W. Y., and Mao, B. H. (2005). "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing. Lecture Notes in Computer Science Vol. 3644*, eds D. S. Huang, X. P. Zhang, G. B. Huang (Berlin: Springer). doi: 10.1007/11538059_91

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). "ADASYN: adaptive synthetic sampling approach for imbalanced learning," in *International Joint Conference on Neural Networks* (Nashville, TN: IEEE World Congress on Computational Intelligence), 1322–1328.

He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi: 10.1109/TKDE.2008.239

Jin, S. H., and Chung, C. K. (2017). Electrophysiological resting-state biomarker for diagnosing mesial temporal lobe epilepsy with hippocampal sclerosis. *Epilepsy Res.* 129, 138–145. doi: 10.1016/j.eplepsyres.2016.11.018

Kassahun, Y., Perrone, R., De Momi, E., Berghöfer, E., Tassi, L., Canevini, M. P., et al. (2014). Automatic classification of epilepsy types using ontology-based and genetics-based machine learning. *Artif. Intell. Med.* 61, 79–88. doi: 10.1016/j.artmed.2014.03.001

Kaur, P., Bharti, V., and Maji, S. (2020). Enhanced epileptic seizure detection using imbalanced classification. *Int. J. Recent Technol. Eng.* 9, 2412–2420. doi: 10.35940/ijrte.A2894.059120

Khambhati, A. N., Bassett, D. S., Oommen, B. S., Chen, S. H., Lucas, T. H., Davis, K. A., et al. (2017). Recurring functional interactions predict network architecture of interictal and ictal states in neocortical epilepsy. *eNeuro* 8:ENEURO.0091-16.2017. doi: 10.1523/ENEURO.0091-16.2017

Kiral-Kornek, I., Roy, S., Nurse, E., Mashford, B., Karoly, P., Carroll, T., et al. (2018). Epileptic seizure prediction using big data and deep learning: toward a mobile system. *EBioMedicine* 27, 103–111. doi: 10.1016/j.ebiom.2017.11.032

Krawczyk, B., Wozniak, M., and Schaefer, G. (2014). Cost-sensitive decision tree ensembles for ef ective imbalanced classification. *Appl. Soft Comput.* 14, 554–562. doi: 10.1016/j.asoc.2013.08.014

Kubat, M., and Matwin, S. (1997). "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the 14th International Conference on Machine Learning*, 179–186

Kuncheva, L. I. (2014). *Combining Pattern Classifiers: Methods and Algorithms, 2nd Edn.* John Wiley & Sons. doi: 10.1002/9781118914564

Lagarde, S., Roehri, N., Lambert, I., Trebuchon, A., McGonigal, A., Carron, R., et al. (2018). Interictal stereotactic-EEG functional connectivity in refractory focal epilepsies. *Brain* 141, 2966–2980 doi: 10.1093/brain/awy214

Laurikkala, J. (2001). "Improving Identification of Difficult Small Classes by Balancing Class Distribution," in *Artificial Intelligence in Medicine. AIME 2001. Lecture Notes in Computer Science, Vol. 2101*, eds S. Quaglini, P. Barahona, and S. Andreassen (Berlin; Heidelberg: Springer), 63–66. doi: 10.1007/3-540-48229-6_9

Lee, P. H. (2014). Resampling methods improve the predictive power of modeling in class-imbalanced datasets. *Int. J. Environ. Res. Public Health*. 11, 9776–9789. doi: 10.3390/ijerph110909776

Lopes da Silva, F., Pijn, J. P., and Boeijinga, P. (1989). Interdependence of EEG signals: linear vs. nonlinear associations and the significance of time delays and phase shifts. *Brain Topogr.* 2, 9–18. doi: 10.1007/BF01128839

López, V., Fernández, A., García, S., Palade, V., and Herrera, F. (2013). An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* 250, 113–141. doi: 10.1016/j.ins.2013.07.007

Loyola-González, O., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., and García-Borroto, M. (2016). Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases. *Neurocomputing* 175, 935–947. doi: 10.1016/j.neucom.2015.04.120

Lüders, H. O., Najm, I., Nair, D., Widdess-Walsh, P., and Bingman, W. (2006). The epileptogenic zone: general principles. *Epilept. Disord.* 8 (Suppl. 2):S1–9.

Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognit.* 91, 216–223. doi: 10.1016/j.patcog.2019.02.023

Masum, M., Shahriar, H., and Haddad, H. M. (2020). "Epileptic seizure detection for imbalanced datasets using an integrated machine learning approach," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (Montréal, QC), 5416–5419. doi: 10.1109/EMBC44109.2020.9175632

Mena, L., and Gonzalez, J. A. (2006). "Machine learning for imbalanced datasets: application in medical diagnostic," in *FLAIRS 2006–Proceedings of the Nineteenth International Florida Artificial Intelligence Research Society Conference* (Melbourne Beach, FL), 574–579.

Narasimhan, S., Kundassery, K. B., Gupta, K., Johnson, G. W., Wills, K. E., Goodale, S. E., et al. (2020). Seizure-onset regions demonstrate high inward directed connectivity during resting-state: an SEEG study in focal epilepsy. *Epilepsia* 61, 2534–2544. doi: 10.1111/epi.16686

Oldham, S., Fulcher, B., Parkes, L., Arnatkeviciute, A., Suo, C., and Fornito, A. (2019). Consistency and differences between centrality measures across distinct classes of networks. *PLoS ONE* 14:e0220061. doi: 10.1371/journal.pone.0220061

Olejarczyk, E., Marzetti, L., Pizzella, V., and Zappasodi, F. (2017). Comparison of connectivity analyses for resting state EEG data. *J. Neural Eng.* 14, 1–13. doi: 10.1088/1741-2552/aa6401

Peter, J., Khosravi, M., Werner, T. J., and Alavi, A. (2018). Global temporal lobe asymmetry as a semi-quantitative imaging biomarker for temporal lobe epilepsy lateralization: a machine learning classification study. *Hell. J. Nucl. Med.* 21, 95–101. doi: 10.1967/s002449910800

Rokach, L. (2010). Ensemble-based classifiers. *Artif. Intell. Rev.* 33, 1–39. doi: 10.1007/s10462-009-9124-7

Roland, J. L., Griffin, N., Hacker, C. D., Vellimana, A. K., Akbari, S. H., Shimony, J. S., et al. (2017). Resting-state functional magnetic resonance imaging for surgical planning in pediatric patients: a preliminary experience. *J. Neurosurg. Pediatr.* 20, 583–590. doi: 10.3171/2017.6.PEDS1711

Romaissa, D., Habib, M., and Chikh, M. A. (2019). "Epileptic seizure detection from imbalanced EEG signal," in *2019 International Conference on Advanced Electrical Engineering, ICAEE 2019* (Macau), 1–6. doi: 10.1109/ICAEE47123.2019.9015113

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Ryvlin, P., Cross, J. H., and Rheims, S. (2014). Epilepsy surgery in children and adults. *Lancet Neurol.* 13, 1114–1126. doi: 10.1016/S1474-4422(14)70156-5

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., and Napolitano, A. (2010). RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man, Cybern. Part A Syst. Hum.* 40, 185–197. doi: 10.1109/TSMCA.2009.2029559

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *J. Am. Stat. Assoc.* 81, 826–831. doi: 10.1080/01621459.1986.10478341

Silfverhuth, M. J., Hintsala, H., Kortelainen, J., and Seppanen, T. (2012). Experimental comparison of connectivity measures with simulated EEG signals. *Med. Biol. Eng. Comput.* 50, 683–688. doi: 10.1007/s11517-012-0911-y

Soriano, M. C., Niso, G., Clements, J., Ortín, S., Carrasco, S., Gudín, M., et al. (2017). Automated detection of epileptic biomarkers in resting-state interictal MEG data. *Front. Neuroinform.* 11:43. doi: 10.3389/fninf.2017.00043

Spencer, S., and Huh, L. (2008). Outcomes of epilepsy surgery in adults and children. *Lancet Neurol.* 7, 525–537. doi: 10.1016/S1474-4422(08)70109-1

Stefanowski, J., and Wilk, S. (2008). "Selective pre-processing of imbalanced data for improving classification performance," in *Data Warehousing and Knowledge Discovery Lecture Notes in Computer Science*, eds I. Y. Song, J. Eder, T. M. Nguyen (Berlin; Heidelberg: Springer), 5182. doi: 10.1007/978-3-540-85836-2_27

Sun, Y., Kamel, M. S., Wong, A. K. C., and Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit.* 12, 3358–3378. doi: 10.1016/j.patcog.2007.04.009

Talairach, J., and Bancaud, J. (1966). Lesion, "irritative" zone and epileptogenic focus. *Confin. Neurol.* 27, 91–94. doi: 10.1159/000103937

Tang, S., and Chen, S. P. (2008). "The generation mechanism of synthetic minority class examples," in *International Conference on Information Technology and Applications in Biomedicine* (Shenzhen), 444–447. doi: 10.1109/ITAB.2008.4570642

Varotto, G., Tassi, L., Franceschetti, S., Spreafico, R., and Panzica, F. (2012). Epileptogenic networks of type II focal cortical dysplasia: a stereo-EEG study. *Neuroimage* 61, 591–598. doi: 10.1016/j.neuroimage.2012.03.090

Varotto, G., Tassi, L., Rotondi, F., Spreafico, R., Franceschetti, S., and Panzica, F. (2013). "Effective brain connectivity from intracranial eeg recordings: identification of epileptogenic zone in human focal epilepsies," in *Modern Electroencephalographic Assessment Techniques*, ed V. Sakkalis (New York, NY: Humana Press). doi: 10.1007/7657_2013_61

Vlachos, I., Krishnan, B., Treiman, D. M., Tsakalis, K., Kugiumtzis, D., and Iasemidis, L. D. (2017). The concept of effective inflow: application to interictal localization of the epileptogenic focus from iEEG. *IEEE Trans. Biomed. Eng.* 64, 2241–2252. doi: 10.1109/TBME.2016.2633200

Wang, S., and Yao, X. (2009). "Diversity analysis on imbalanced data sets by using ensemble models," in *2009 IEEE Symposium on Computational Intelligence and Data Mining*, 324–331. doi: 10.1109/CIDM.2009.4938667

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorat. Newslett.* 6, 7–19. doi: 10.1145/1007730.1007734

Wendling, F., Ansari-Asl, K., Bartolomei, F., and Senhadji, L. (2009). From EEG signals to brain connectivity: a model-based evaluation of interdependence measures. *J. Neurosci. Methods* 183, 9–18. doi: 10.1016/j.jneumeth.2009.04.021

Wendling, F., Chauvel, P., Biraben, A., and Bartolomei, F. (2010). From intracerebral EEG signals to brain connectivity: identification of epileptogenic networks in partial epilepsy. *Front. Syst. Neurosci.* 4:154. doi: 10.3389/fnsys.2010.00154

Xie, C., Du, R., Ho, J. W., Pang, H. H., Chiu, K. W., Lee, E. Y., et al. (2020). Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. *Eur. J. Nucl. Med. Mol. Imaging* 47, 2826–2835. doi: 10.1007/s00259-020-04756-4

Yen, S. J., and Lee, Y. S. (2006). "Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset," in *Intelligent Control and Automation. Lecture Notes in Control and Information Sciences,* eds D. S. Huang, K. Li, G. W. Irwin (Kunming: Springer). doi: 10.1007/978-3-540-37256-1_89

Zhou, Z. H., and Liu, X. Y. (2006). "Training cost-sensitive neural networks with methods addressing the class imbalance problem," in *IEEE Transactions on Knowledge and Data Engineering*, 63–77. doi: 10.1109/TKDE.2006.17

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 24 November 2021
doi: 10.3389/fninf.2021.742807

Multimodal Ensemble Deep Learning to Predict Disruptive Behavior Disorders in Children

*Sreevalsan S. Menon and K. Krishnamurthy**

Department of Mechanical and Aerospace Engineering, Missouri University of Science and Technology, Rolla, MO, United States

Oppositional defiant disorder and conduct disorder, collectively referred to as disruptive behavior disorders (DBDs), are prevalent psychiatric disorders in children. Early diagnosis of DBDs is crucial because they can increase the risks of other mental health and substance use disorders without appropriate psychosocial interventions and treatment. However, diagnosing DBDs is challenging as they are often comorbid with other disorders, such as attention-deficit/hyperactivity disorder, anxiety, and depression. In this study, a multimodal ensemble three-dimensional convolutional neural network (3D CNN) deep learning model was used to classify children with DBDs and typically developing children. The study participants included 419 females and 681 males, aged 108–131 months who were enrolled in the Adolescent Brain Cognitive Development Study. Children were grouped based on the presence of DBDs ($n = 550$) and typically developing ($n = 550$); assessments were based on the scores from the Child Behavior Checklist and on the Schedule for Affective Disorders and Schizophrenia for School-age Children-Present and Lifetime version for DSM-5. The diffusion, structural, and resting-state functional magnetic resonance imaging (rs-fMRI) data were used as input data to the 3D CNN. The model achieved 72% accuracy in classifying children with DBDs with 70% sensitivity, 72% specificity, and an F1-score of 70. In addition, the discriminative power of the classifier was investigated by identifying the cortical and subcortical regions primarily involved in the prediction of DBDs using a gradient-weighted class activation mapping method. The classification results were compared with those obtained using the three neuroimaging modalities individually, and a connectome-based graph CNN and a multi-scale recurrent neural network using only the rs-fMRI data.

Keywords: deep learning, disruptive behavior disorders, multimodal ensemble learning, neuroimaging, 3D CNN

OPEN ACCESS

Edited by:
Itir Onal Ertugrul,
Tilburg University, Netherlands

Reviewed by:
Fahad Saeed,
Florida International University, United States
Rajat Mani Thomas,
Academic Medical Center, Netherlands
Yunyan Zhang,
University of Calgary, Canada

****Correspondence:***
K. Krishnamurthy
kkrishna@mst.edu

Received: *16 July 2021*
Accepted: *15 October 2021*
Published: *24 November 2021*

Citation:
Menon SS and Krishnamurthy K (2021) Multimodal Ensemble Deep Learning to Predict Disruptive Behavior Disorders in Children. Front. Neuroinform. 15:742807. doi: 10.3389/fninf.2021.742807

1. INTRODUCTION

Magnetic resonance imaging (MRI) is a powerful noninvasive neuroimaging tool that can reveal anatomical features and neuronal activities inside a brain. MRI data is widely used to study cognitive development, pathologies, and psychiatric disorders. Diffusion MRI (dMRI) can reveal information about the microstructures, fiber connections, and anatomical connectivities within the brain, and the static anatomical images acquired using structural MRI (sMRI) provide information about the gross anatomical structures in the brain. Dynamic activities inside the brain are measured using functional MRI (fMRI), which is used to identify brain activities in the absence of a task (resting-state fMRI; rs-fMRI) or during a task (task fMRI; tfMRI).

Disruptive behavior disorders (DBDs) include oppositional defiant disorder (ODD; a pattern of angry/irritable mood, argumentative/defiant behavior, or vindictiveness lasting at least 6 months) and conduct disorder (CD; behavior in which the basic rights of others or major age-appropriate societal norms or rules are violated; American Psychiatric Association, 2013). They are prevalent in children and the most common reasons for referring children to mental health services (Hawes et al., 2020). ODD is estimated to occur in 2–16% of youth, depending on the population being studied and the method for diagnosis, and CD, which is more prevalent among younger males, rates range from 6 to 9% (SAMHSA, 2011). DBDs are associated with increased risk for other mental health and substance use disorders (Nock et al., 2006), and are predictors of poor mental health conditions (Scarmeas et al., 2007). These disorders can cause substantial economic losses for society in terms of service utilization (Rivenbark et al., 2018). Therefore, early diagnosis of DBDs is crucial to lower the risk for subsequent disorders with appropriate psychosocial interventions and treatment. However, DBDs are challenging to diagnose as they are often comorbid with other disorders, such as attention-deficit/hyperactivity disorder, anxiety, and depression (Allen et al., 2020).

Machine learning concepts are now receiving increased attention for analysis and prediction in neuroimaging applications. Traditional machine learning techniques require hand-engineered feature selection, which are time-consuming and prone to bias due to manual feature selection. Deep learning is a recent development in machine learning that overcomes the issues associated with hand-engineering and requisite domain expertise for feature selection. Deep learning is a representation learning in which raw data are fed into a learning algorithm that decomposes it into multiple levels of complex nonlinear representative patterns of the input data (LeCun et al., 2015). The burgeoning wide applications of deep learning models can be attributed to the implementation of a convolutional neural network (CNN) because it cut the second-best error rate for image classification by nearly half at the ImageNet Large-Scale Visual Recognition Challenge in 2012 (Krizhevsky et al., 2012). With the advent of parallel computing and graphics processing units, deep representation learning was successfully implemented in numerous areas, such as image processing and analysis tasks, natural language processing, speech recognition, and data synthesis and analysis (LeCun et al., 2015). A large number of medical image analyses now focus on applying deep learning methods to extract features from raw data for further analysis and interpretation (Lundervold and Lundervold, 2019).

CNNs inspired by visual neuroscience are one of the widely used deep learning architectures. A typical CNN includes a convolutional layer, pooling layer, and fully connected layer. The convolutional layer consists of filters/kernels of fixed size that strides with a partial overlap through the input and generates feature maps that are locally weighted sum of input features. Each filter in a convolutional layer looks for the same pattern in different parts of the input, and outputs a unique feature map. The convolution filter thus looks for highly correlated local motifs that can occur at any location in the input (LeCun et al., 2015). The feature maps in the convolution layer are then passed through nonlinear activation functions, such as the rectified linear unit (ReLu) (O'Shea and Hoydis, 2017). The output from one or more convolution layers is then pooled in a pooling layer that merges similar features. Pooling filters output the average or maximum value inside the filter grid and impart translational invariance, for inputs with minor shifts and distortions in rows or columns, to the activation map. Typically, several convolutional and pooling layers are stacked in a CNN, and they are followed by a fully connected layer. The fully connected layer usually connects to an output layer, which could be a softmax function for classification tasks or a linear or support vector machine for regression tasks. CNNs learn in a hierarchical fashion from low-level features, such as edges (similar to primary visual cortex), to high-level features, such as shapes (identical to the secondary visual cortex), in deep layers similar to the hierarchical structure in a human visual cortex (Hubel and Wiesel, 1962). Brainnet CNN (Kawahara et al., 2017) is an earlier developed connectome-based graph CNN which is composed of edge-to-edge, edge-to-node, and node-to-graph convolutional filters that leverage the topological locality of brain networks as opposed to local spatial filtering.

CNNs are often considered "black boxes" that perform classifications without explanations on what a model learned or which part of an input was responsible for the classification. One primary goal of machine learning in neuroimaging is to reveal neuromarkers that are indicative of brain health, and diseases and disorders (Khosla et al., 2019). To address these issues, visualization techniques can be utilized to discover discriminative features learned by a CNN model. Class activation mapping (CAM) is a technique to obtain visual explanations of the input regions that a CNN emphasized in its classification (Zhou et al., 2016; Selvaraju et al., 2017) by calculating the derivative of the CNN classification function estimated via back-propagation with respect to the input data. Gradient CAM (Grad-CAM) and Grad-CAM++ are two improved versions of CAMs because they can be applied to a wide variety of networks without global average pooling and retraining, and they reveal the discriminative regions in any CNN architecture (Selvaraju et al., 2017; Chattopadhyay et al., 2018). The three CAM techniques were compared in one study on classifying multiple sclerosis types, and it was shown that Grad-CAM outperformed CAM and Grad-CAM++ (Zhang et al., 2021).

A multi-scale recurrent neural network (MsRNN) is another deep learning-based framework that can directly work on the dynamic spatiotemporal fluctuations in the brain activity measured using rs-fMRI time courses for identifying brain disorders (Yan et al., 2019). While the CNN models, *deep in space*, can be used as an "encoder" for obtaining correlations between brain regions, recurrent neural network (RNN) models, *deep in time*, can be utilized in sequence classification (Yan et al., 2019). A simple RNN consists of input, hidden and output layers, and it processes the input sequentially with respect to time. The distinguishing feature of RNNs is that the output from a layer is used as input for the layer itself, thereby forming a feedback loop. This allows the RNN to have a history of the sequence elements that can be used to predict the upcoming sequence elements.

Several studies in machine learning showed that the performance of the learning algorithm can be improved using ensemble learning, which is an algorithm-independent machine learning strategy (Opelt et al., 2004; Khosla et al., 2019). Moreover, brain abnormalities are heterogeneous and cause alterations in functional connectivity and structural changes (McLaughlin et al., 2019). Studies have found abnormal brain activities in children with DBDs using dMRI (Hummer et al., 2015), sMRI (Wallace et al., 2014; Hummer et al., 2015; Waller et al., 2020), tfMRI (Rubia et al., 2009; Hawes et al., 2020), and rs-fMRI (Lu et al., 2015; Werhahn et al., 2020). Therefore, there is significant motivation to take advantage of complementary information on various aspects of neuropathology. This study addresses a knowledge gap in the availability of multimodal tools for studying brain abnormalities using different neuroimaging modalities.

In this study, a 3D CNN ensemble deep learning model framework with multimodal neuroimaging data was exploited to identify children with DBDs. The dMRI, sMRI, and rs-fMRI data from a subsample of children enrolled in the Adolescent Brain Cognitive Development (ABCD) Study (Casey et al., 2018) were used as the input data. Furthermore, the brain regions involved in classifying children with DBDs were identified utilizing Grad-CAM that illustrated the discrimination power of the classifier and the ability to identify neuroimaging phenotypes for DBDs. To assess improvements offered by the ensemble learning, the results were compared with those obtained using the three neuroimaging modalities individually; they were also compared with those obtained using two other readily available deep-learning frameworks, Brainnet CNN and an MsRNN, model with rs-fMRI data. We hypothesized that the classification performance of the ensemble deep learning model will be significantly better than the single modality models.

2. MATERIALS AND METHODS

2.1. Dataset

Data used in this study came from the ABCD Study that recruited 11,878 children (48% female; 52% male) between 108 and 120 months of age across 21 sites in the United States. A detailed description of the recruitment, demographics, physical health, and mental assessment and imaging protocols for the study can be found elsewhere (Barch et al., 2018; Casey et al., 2018; Garavan et al., 2018). The baseline ABCD Study data used in this study were from the annual 2.0.1 data release and can be downloaded from the National Institute of Mental Health (NIMH) Data Archive[1]. The data is available to qualified researchers at no cost after their NIMH Data Archive Data Use Certification has been approved. Children with DBDs were identified using the Child Behavior Checklist (CBCL) and the Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime version for DSM-5 (K-SADS-PL) (Hawes et al., 2020). Specifically, the criterion included children who: (i) scored at or above the borderline clinical range (i.e., T-scores $\geq$67) on either the CBCL DSM-oriented conduct problems subscale or oppositional defiant problems subscale; or (ii) received a K-SADS-PL conduct disorder or oppositional defiant disorder diagnosis. Based on this criterion, there were 1,100 children with minimally preprocessed data with all three neuroimaging modalities, i.e., dMRI, sMRI, and rs-fMRI.

2.2. Preprocessing of ABCD Study Minimally Preprocessed Data

DTI data were preprocessed using FSL (FMRIB's Software Library[2]) scripts, which were used to perform nonlinear registration and projection onto an alignment-invariant tract representation of fractional anisotropy (FA) and mean diffusivity (MD). First, diffusion tensor models were fit at each voxel by using FMRIB's Diffusion Toolbox (FDT, part of FSL). Second, brain extraction was performed using the brain extraction tool (BET) (Smith, 2002). Third, nonlinear registration was done, thereby aligning all FA and MD images to a FMRIB58_FA standard-space image, which has a $1 \times 1 \times 1$ mm resolution, as the target. Finally, all images were resampled back to the $2 \times 2 \times 2$ mm FSL default MNI152 standard-space template resolution. **Figure S1** shows an example DTI image.

The sMRI T1-weighted images were preprocessed mainly using the FSL software. First, extraction of the brain tissue from the skull was performed by using BET. Second, registration to standard space images was carried out using FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002). Third, registration from high-resolution structural to the FSL default MNI152 standard space was then further refined using FNIRT nonlinear registration (Andersson et al., 2007a,b). Finally, the FMRIB's Automated Segmentation Tool (FAST) (Zhang, 2001) was used to segment the brain 3D-image into three different tissue types: (i) gray matter; (ii) white matter; and (iii) cerebrospinal fluid (CSF). **Figure S2** shows an example sMRI image.

The rs-fMRI data preprocessing was carried out using FEAT (FMRI Expert Analysis Tool) Version 6.00, a part of FSL. Registration to high-resolution structural and the FSL default MNI152 standard space images was carried out using FLIRT. Registration from high-resolution structural to standard-space was further refined using FNIRT nonlinear registration. Additionally, the following pre-statistics processing was applied: (i) motion correction using MCFLIRT (Jenkinson et al., 2002); (ii) non-brain removal using BET; (iii) spatial smoothing using a Gaussian kernel of FWHM 8.0 mm; (iv) grand-mean intensity normalization of the entire 4D dataset by a single multiplicative factor, which was done by default in all the fMRI software packages to ensure each image scan had roughly the same mean; and (v) high-pass temporal filtering (Gaussian-weighted least-squares straight-line fitting, with sigma = 50.0 s). The Pearson seed-based correlation values were calculated for four regions of interest, namely posterior and anterior cingulate cortex (PCC and ACC), medial prefrontal cortex (mPFC) and ventral caudate, which are known to be affected in children with DBDs (Alegria et al., 2016). **Figure S3** shows an example rs-fMRI image for the ACC.

[1]https://dx.doi.org/10.15154/1504041

[2]www.fmrib.ox.ac.uk/fsl

TABLE 1 | Demographic and clinical characteristics of the study pool.

Characteristic	DBDs			TD			*p*-value
	Mean	SD	%	Mean	SD	%	
Demographics							
Age (months)	118.3	7.7		118.7	7.4		0.38
Sex (male)			61.6			62.2	0.85
Race							
African American			16.2			14.0	0.44
Caucasian			54.0			53.8	
Hispanic			16.2			19.5	
Other			13.6			12.7	
Clinical							
CBCL CP subscale	63.6	8.13		50	0		<0.001
CBCL ODD subscale	63.9	7.44		50	0		<0.001
KSADS-PL CD diagnosis			29.6			0	<0.001
KSADS-PL ODD diagnosis			73.3			0	<0.001

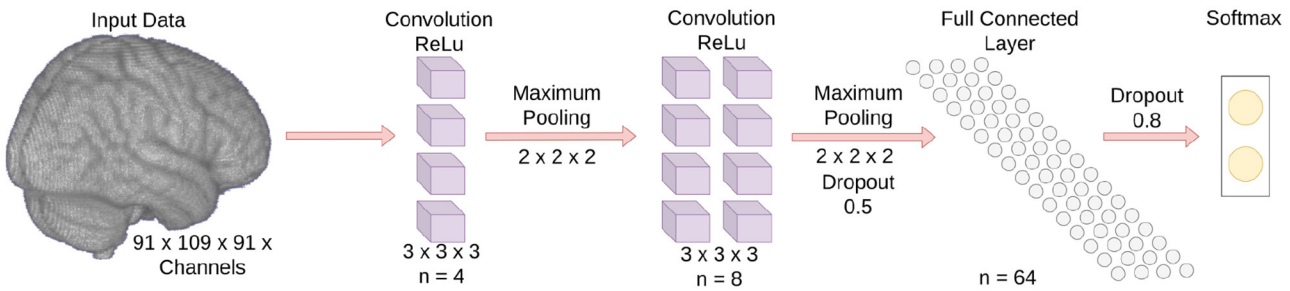


FIGURE 1 | Schematic of the 3D CNN model. The number of input data channels is not shown because it varies depending on the input modality.

Children were removed from the study pool following preprocessing due to high motion (framewise displacement > 0.25 mm), misalignment, and registration failures. As a result, the complete preprocessed data were available for 550 children and a matching number of children in age and sex without DBDs (typically developing, TD) were selected from the ABCD Study data as the control group. **Table 1** shows the demographic and clinical characteristics of the final study pool. Descriptive statistics show that the groups were equivalent on demographic variables and significantly different on clinical scores.

2.3. Ensemble Learning

Three multichannel 3D CNNs whose inputs were dMRI, sMRI, and rs-fMRI, respectively, were trained in this study to classify children with DBDs and TD children. The goal for the 3D CNNs was to learn the mapping between input (features related to the microstructural integrity and gross anatomical structure of the brain, and resting-state functional patterns) and label (TD children and children with DBDs), so that the 3D CNNs can predict DBDs in previously unseen children. As shown in **Figure 1**, each 3D CNN model had two convolution blocks each consisting of a 3D convolutional layer (kernel size 3, stride 1), a ReLU activation layer, and a max-pooling layer (kernel size 2, stride 2). The number of feature channels were 4 and 8 for the convolution layers, respectively. The last layer was a fully connected layer with 64 neurons to combine the feature vectors, and a dropout layer was used to reduce model overfitting. The output was a softmax classification layer. The input channels for the three 3D CNN models were as follows: (i) dMRI model—two channels for FA and MD values; (ii) sMRI model—three channels for gray matter, white matter, and CSF; and (iii) rs-fMRI model—four channels for Pearson correlation of seed regions ACC, PCC, mPFC, and ventral caudate. The three models were combined in an ensemble learning strategy that gave equal weight during maximum voting of the softmax output for classifying children with DBDs and TD children.

The 3D CNN models were trained with mini-batch sizes of 32 with early stopping conditioned on validation accuracy. The binary cross-entropy was used as the loss function and the neural network weights were optimized using the Adam optimizer. The learning rate and gradient decay were set to 0.001 and 0.9, respectively. The squared gradient decay, epsilon, and maximum epochs were set to 0.9, 0.001, and 50, respectively. No attempt was made to optimize the aforementioned parameters.

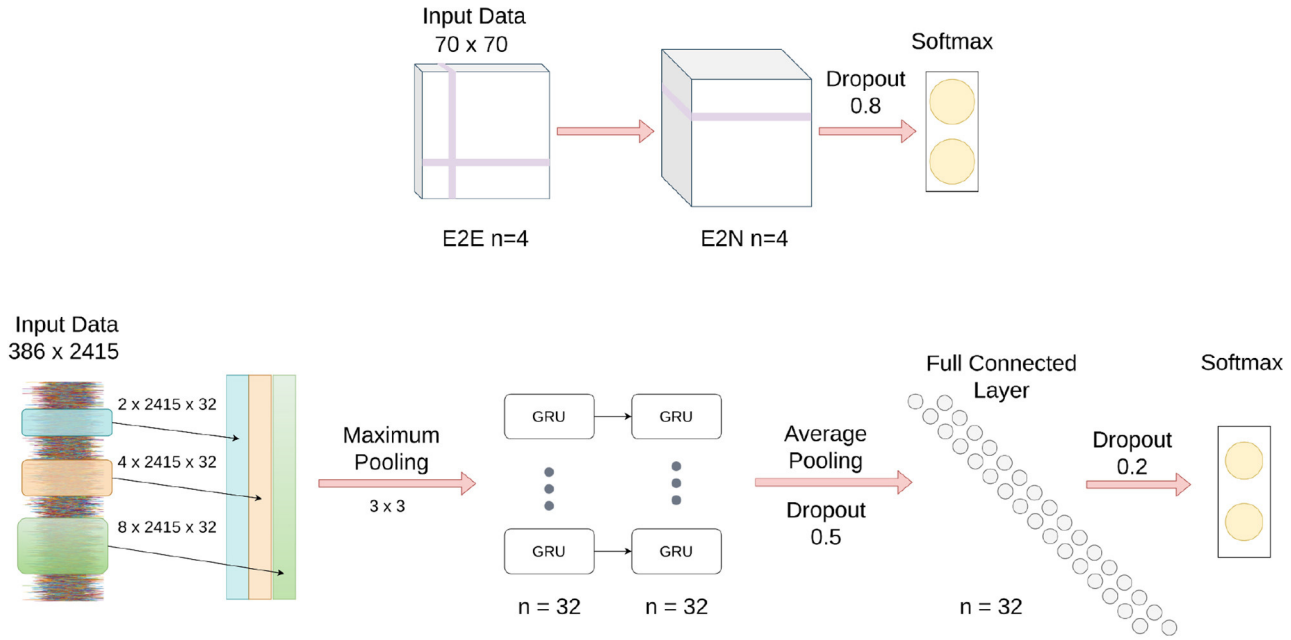


FIGURE 2 | Schematic of Brainnet CNN and MsRNN architecture. **(Top)** Brainnet CNN; the edge to edge (E2E) layer uses a crosshair convolution filter, and the edge to node (E2N) layer uses a 1D convolution row filter. **(Bottom)** MsRNN; three varied convolutions are first performed in the input layer, the output is then concatenated, and finally maximum pooled before being fed into the gated recurrent units (GRUs).

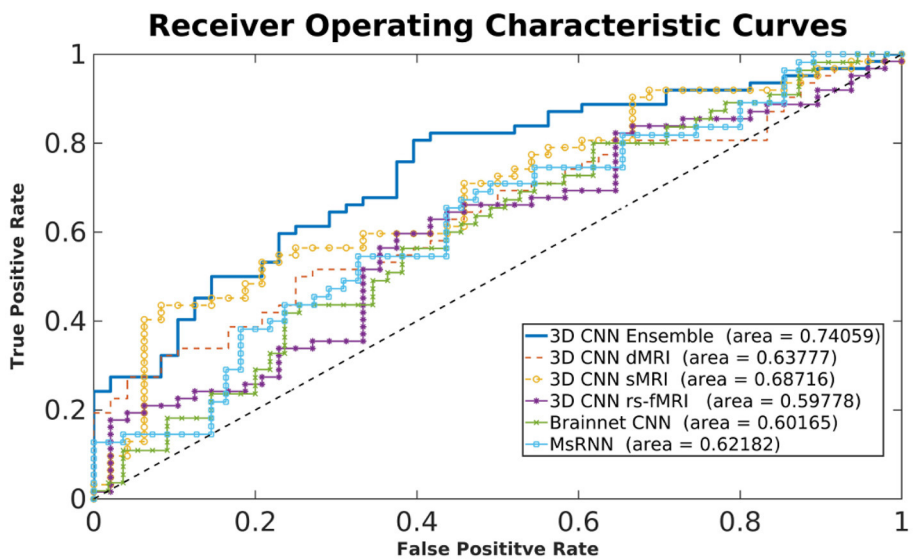


FIGURE 3 | Typical receiver operating characteristic curves for different models.

To ensure that all 3D CNN models relied on information from the same voxels, the FSL default MNI152 standard-space mask was applied to the voxel-level data before feeding into the 3D CNN model (Khosla et al., 2019). This step removed voxels that may have emerged outside the standard brain template because the preprocessing transformation matrix does not create the exact brain boundary.

2.4. Brainnet CNN

As shown in **Figure 2**, input to the Brainnet CNN was a functional connectivity matrix obtained using timeseries extracted from 70 resting-state networks, which were identified using publicly available 70-component independent component analysis maps (Smith et al., 2009). The blood oxygenation level-dependent (BOLD) timeseries were extracted from the 70 brain

TABLE 2 | Classification performance in percentage.

Method	Modality	Accuracy			Sensitivity	Specificity	F1-score
		Mean (SD)	*p*-value	Cohen's *d*	Mean (SD)	Mean (SD)	Mean (SD)
3D CNN Ensemble		72 (4.5)	Proposed model		70 (17.0)	72 (15.6)	70 (9.0)
3D CNN	dMRI	64 (2.6)	<0.001	2.20	60 (16.0)	67 (14.3)	61 (9.7)
	sMRI	66 (2.2)	<0.001	1.85	64 (11.2)	65 (13.2)	64 (6.4)
	rs-fMRI	66 (3.0)	0.002	1.57	62 (15.4)	69 (16.4)	64 (7.2)
BrainnetCNN	rs-fMRI	62 (2.9)	<0.001	2.67	60 (7.3)	64 (4.3)	61 (4.5)
MsRNN	rs-fMRI	62 (2.5)	<0.001	2.79	56 (7.7)	68 (8.0)	59 (4.4)

areas by averaging the BOLD signal over all voxels belonging to each brain area. The timeseries were detrended and demeaned, and the data were bandpass filtered in the range of 0.01–0.15 Hz to improve identification of the resting-state fluctuations (Menon and Krishnamurthy, 2019a). The functional connectivity matrix was obtained using Pearson correlation with normalization to z-scores using the Fisher transformation.

The Brainnet CNN model was implemented in Python by modifying publicly available scripts (Kawahara et al., 2017). The Brainnet CNN model had an edge-to-edge (E2E) layer with four filters, followed by a edge-to-node (E2N) layer with four filters, and finally a dense layer with two neurons. A leaky ReLU non-linearity with alpha equal to 0.33 was applied to the output of each layer except the last layer, which was a softmax layer. Dropout regularization with a rate of 0.8 was used for the edge-to-node layer and cross-entropy loss was used to optimize the classification model. The models were trained for 1,000 iterations using stochastic gradient descent with a momentum equal to 0.9. The learning rate was set to 0.001 and a decay of 0.0005 was used for the classification model. No attempt was made to optimize the aforementioned parameters.

2.5. Multi-Scale Recurrent Neural Network

Figure 2 shows a schematic of the MsRNN used in this study. The timeseries extracted from 70 resting-state networks that were input into the Brainnet CNN were also used as the input to an MsRNN. The dynamic correlation connectivity values of 2,415 edges were calculated with a window length of 85 TR and step size of 5 TR (Menon and Krishnamurthy, 2019a). The MsRNN utilized three different scales of 32 1D convolutional filters (2 TR, 4 TR, and 8 TR, TR = 0.8 s), one concatenation layer, one max-pooling layer of kernel size 3, a two-layer stacked gated recurrent unit GRU with 32 filters which were densely connected in a feed-forward manner, and an averaged layer that integrated the whole sequence followed by a dense layer of 32 neurons before the softmax classification layer. Dropout layers were used before and after the dense neurons with 50 and 20% dropout, respectively, and L1 and L2 regularization of 0.01 was used to avoid overfitting the data. The MsRNN was trained in Python following Yan et al. (2019) with a mini-batch size of 32, and included early stopping conditioned on validation accuracy and a learning rate of 0.001. The binary cross-entropy was used as the loss function, and the neural network weights were optimized using the Adam optimizer. No attempt was made to optimize the aforementioned parameters.

3. RESULTS

3.1. Experiments

To test the efficacy of the multimodal data ensemble, a ten-fold cross-validation (CV) strategy with maximum voting was investigated. The Grad-CAM method was applied to the predicted output, and the results for all the children with DBDs and TD children were averaged to delineate the global trends of the important regions involved in the classification. To benchmark the performance of the ensemble learning approach, the results were compared to those obtained from: (i) the three 3D CNN models used in the ensemble learning considered individually; (ii) Brainnet CNN; and (iii) MsRNN model.

3.2. Classification Performance

Figure 3 shows typical receiver operating characteristic curves and **Table 2** shows the performance of the different methods for classifying children with DBDs and TD controls. With 10-fold cross-validation, the multimodal ensemble model with maximum voting resulted in an average prediction accuracy to 72%. The average prediction accuracies for dMRI, sMRI, and rs-fMRI single modalities were 64, 66, and 66%, respectively, compared to 62% with the Brainnet CNN and MsRNN. **Table 2** also shows the multimodal ensemble model to have higher sensitivity, specificity, and F1-score compared to the other models considered.

Statistical results from two-sample t-tests were used to compare the accuracy of the classification performance of the different models. The higher accuracy of the proposed ensemble model compared to all the other models was significant (highest *p*-value was 0.002) with a very large to huge effect size calculated as Cohen's *d* (Sawilowsky, 2009). Overall, as hypothesized, the classification performance was significantly higher using the ensemble learning model because it utilized complementary information from the three different modalities. The results also indicated the superiority of voxel-based 3D CNN models compared to network-level models, such as Brainnet CNN and MsRNN.

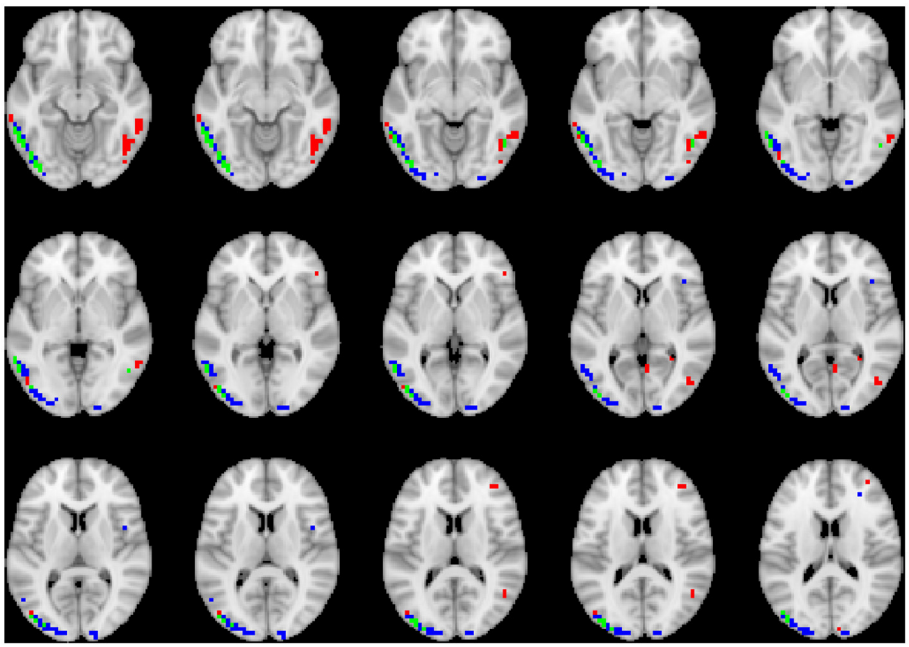


FIGURE 4 | Axial views of voxels primarily contributing to children classification in dMRI image. Green, common to DBD and TD groups; red, DBD group; blue, TD group. The right side of each image corresponds with the right side of the brain.

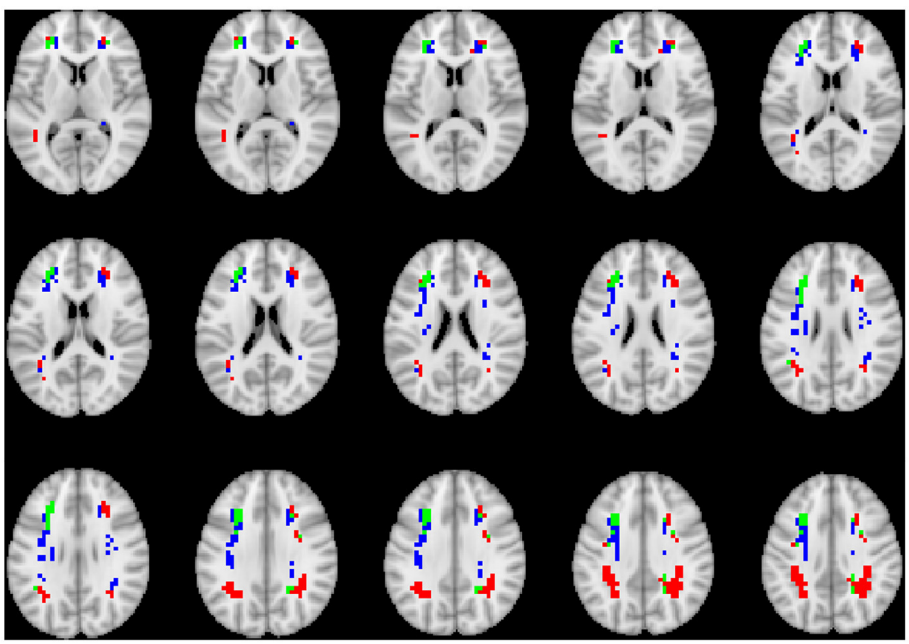


FIGURE 5 | Axial views of voxels primarily contributing to children classification in sMRI image. Green, common to DBD and TD groups; red, DBD group; blue, TD group. The right side of each image corresponds with the right side of the brain.

3.3. Visualization

To visualize the brain regions that primarily contributed to children classification, Grad-CAM obtained for children with DBDs and TD children were thresholded at 99 percentile to first identify voxels with high gradient values. The brain regions that primarily contributed toward classification were then identified using the JHU ICBM-DTI-81 white-matter atlas for dMRI image and the AAL atlas for sMRI and rs-fMRI images. **Figures 4–6**

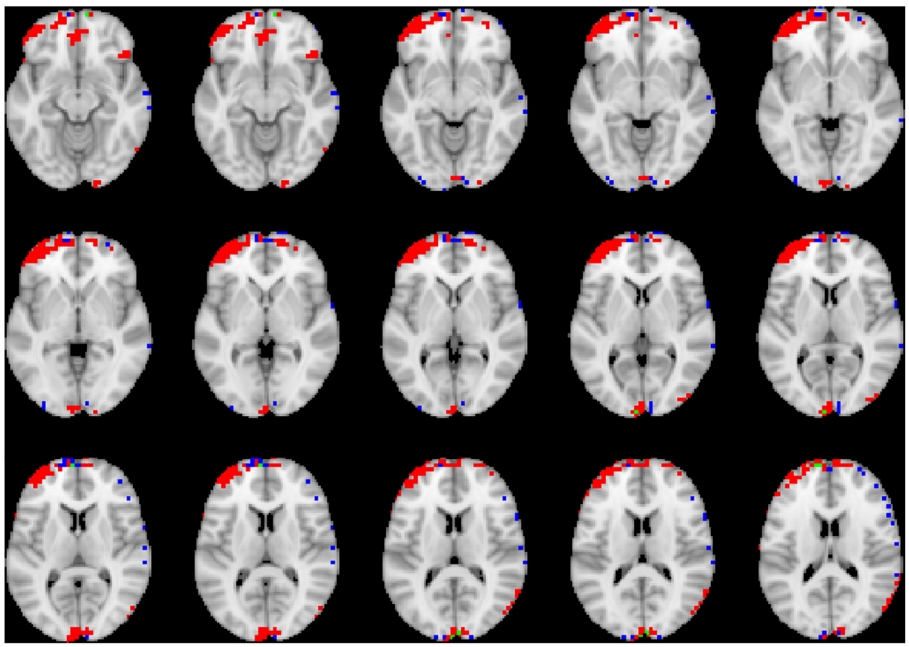


FIGURE 6 | Axial views of voxels primarily contributing to children classification in rs-fMRI image. Green, common to DBD and TD groups; red, DBD group; blue, TD group. The right side of each image corresponds with the right side of the brain.

show axial views of voxels that primarily contributed to the classification of children with DBDs and TD children in dMRI, sMRI, and rs-fMRI images, respectively. **Table 3** lists the top five brain regions in the dMRI images and top ten in the sMRI and rs-fMRI images of children with DBDs and TD children. Some of the regions were common to both groups and are listed in green. On the other hand, regions that were unique to children with DBDs and TD children are listed in red and blue, respectively. These unique regions are of interest because their contributions outweigh the gradient contributions of common regions, hence are highly class discriminative. Further, these unique regions are evidence of abnormalities in children with DBDs because the primarily contributing regions are all different from those in the TD children.

3.4. 3D CNN Training Information

Time for training the 3D CNN model was similar for the three modalities. It typically took around 150 min per fold on Dell Precision 7,910 and 7,920 Tower workstations with Intel Xeon processors, 128 GB RAM and 1 GB GPU. **Figures S4**, **S5** show a typical training graph and Precision-Recall curves, respectively.

4. DISCUSSION

This was the first neuroimaging study to consider the classification of children with DBDs. This is a challenging problem because DBDs are often comorbid with other disorders, such as attention-deficit/hyperactivity disorder, anxiety, and depression. The multimodal ensemble learning approach for diagnosing DBDs with voxel-based 3D CNN is a novel approach and the accuracy of the ensemble model increased by 6–10% compared to other models. The maximum voting in the ensemble learning method simulates how clinicians typically make decisions. Given that brain abnormalities are heterogeneous, it is naturally advantageous to utilize information from multimodalities. The maximum voting is the simplest and easiest ensemble method that can be applied to 3D CNN models. The maximum voting strategy also ensures that the results are not biased toward any single modality, but will take into account all available information. 3D CNN models, unlike traditional machine learning methods, such as artificial neural networks or support vector machines are well suited to include the spatial relations in the 3D neuroimaging data, which are known to affect brain functioning. Furthermore, traditional machine learning methods will overfit the data and reduce the validation classification accuracy with high-dimensional 3D neuroimaging training data.

Grad-CAM reveals the discriminant regions in the brain that contributed to the classification of children with DBDs. As shown in **Table 3**, most of these regions corroborate with results from past studies on abnormal development in children with DBDs. To mention a few, alterations in the white matter integrity of the left inferior fronto-occipital fasciculus were suggested as a potential biomarker of conduct disorder (Graziano et al., 2021). Similarly, superior longitudinal fasciculus areas were shown to have differences in diffusion measurements that suggested poor maturation of structural connections (Hummer et al., 2015) in children with DBDs. Morphological aberrance of frontoparietal

TABLE 3 | Brain regions primarily contributing to children classification.

dMRI	sMRI	rs-fMRI
Superior longitudinal fasciculus L	Middle frontal gyrus L	Medial superior frontal gyrus L
Superior longitudinal fasciculus L - temporal part	Superior frontal gyrus L	Superior frontal gyrus, dorsolateral L
Inferior longitudinal fasciculus L	Angular gyrus L	Rectus gyrus L (Zhang et al., 2017; Cao et al., 2018)
Superior longitudinal fasciculus R (Haney-Caron et al., 2014; Hummer et al., 2015; Lindner et al., 2016; Sarkar et al., 2016; Puzzo et al., 2018)	Precentral gyrus L	Rectus gyrus R
Superior longitudinal fasciculus R - temporal part	Superior frontal gyrus R	Middle frontal gyrus L (Lu et al., 2015; Zhang et al., 2015; Cao et al., 2018)
Inferior fronto-occipital fasciculus L (Haney-Caron et al., 2014; Lindner et al., 2016; Graziano et al., 2021)	Middle frontal gyrus R (Huebner et al., 2008; Fairchild et al., 2015)	Medial orbital superior frontal gyrus L
Forceps major (Lindner et al., 2016)	Postcentral gyrus L (Hyatt et al., 2012)	Medial superior frontal gyrus R (Zhang et al., 2017; Cao et al., 2018)
	Middle temporal gyrus L (Huebner et al., 2008; Fairchild et al., 2011)	Inferior temporal gyrus L (Zhang et al., 2015; Cao et al., 2018; Werhahn et al., 2020)
	Inferior parietal lobule L (Wallace et al., 2014)	Temporal pole superior temporal gyrus L (Cao et al., 2018)
	Middle occipital gyrus R (Huebner et al., 2008)	Inferior parietal lobule L (Zhang et al., 2015)
	Inferior frontal gyrus, triangular part L (Huebner et al., 2008; Fairchild et al., 2011; Hyatt et al., 2012)	Postcentral gyrus R (Lu et al., 2015; Cao et al., 2018, 2019; Lu F. et al., 2020; Werhahn et al., 2020)
	Inferior frontal gyrus, opercular part L (Huebner et al., 2008; Fairchild et al., 2011; Hyatt et al., 2012)	Supramarginal gyrus R (Zhang et al., 2015, 2017)
	Inferior frontal gyrus, opercular part R (Hyatt et al., 2012; Fairchild et al., 2013)	Precentral gyrus R (Lu F. et al., 2020; Werhahn et al., 2020)
	Precentral gyrus R (Hyatt et al., 2012; Fairchild et al., 2013; Jiang et al., 2015)	Middle temporal gyrus R (Lu et al., 2015; Zhang et al., 2015; Wu et al., 2017)
	Hippocampus R (Waller et al., 2020)	Middle frontal gyrus R (Zhang et al., 2015; Cao et al., 2019)
		Inferior frontal gyrus, opercular part R (Zhang et al., 2015; Cao et al., 2018, 2019)
		Inferior frontal gyrus, triangular part R (Zhang et al., 2015; Cao et al., 2018, 2019)
		Superior frontal gyrus R (Zhang et al., 2017; Cao et al., 2018)

Past studies corroborating with results obtained are shown in parentheses. Green, common to DBD and TD groups; red, DBD group; blue, TD group; L, left hemisphere; R, right hemisphere.

and temporal gyrus areas can lead to disruptive behavior (Huebner et al., 2008; Hyatt et al., 2012; Fairchild et al., 2015) and most of these regions were found to be class discriminative in this study. Functional connectivity alterations have been reported for children with DBDs, and class discriminative regions found using grad-CAM were consistent with many of the reported regions (Lu et al., 2015; Werhahn et al., 2020). Functional connectivity values for higher-order cognitive functional regions such as the middle frontal gyrus and superior frontal gyrus were also found to be class discriminative (Lu et al., 2015).

The 72% average accuracy obtained using the ensemble learning approach is good. Because there are no other studies on classifying children with DBDs to benchmark against, some representative neuroimaging studies using deep learning were reviewed to qualify the multimodal ensemble model performance. El Gazzar et al. (2019) trained a 1D-CNN on a publicly available autism dataset with nearly 2000 participants to classify rs-fMRI images with an accuracy of ∼65%. The accuracy improved to 66% with a 3D CNN (Thomas et al., 2020). Lu H. et al. (2020) obtained an accuracy of 61% by applying multi-kernel fuzzy clustering based on an auto-encoder to classify participants with autism spectrum disorder (ASD) using the Autism Brain Imaging Data Exchange (ABIDE) database (nearly 1,050 participants). Using an ensemble approach on ABIDE data, a classification accuracy of 72.3% was obtained by Khosla et al. (2019). Similar to DBDs, classification of ASD using machine learning methods is also considered challenging because it varies from person-to-person in severity and combination of symptoms. Other studies with a classification accuracy >70% are typically in cases where the sample size is <200 (see Vieira et al., 2017; Zhang et al., 2020 for an overview). The sample size is an

important parameter to consider because a negative relationship between accuracy and sample size has been noted (Pulini et al., 2019).

5. LIMITATIONS AND FUTURE DIRECTIONS

The robustness of the training models could not be determined by using a leave-site-out cross-validation scheme for the ABCD Study data that was collected from 21 sites with optimized and harmonized measures and procedures (Casey et al., 2018). A k-fold cross-validation was used instead because the number of children from each site in the study pool was imbalanced. The number of children with DBDs varied among the different sites, from a low of 3 to a maximum of 113.

This study investigated the superiority of ensemble learning for classifying brain disorders. The sample size used here was relatively large compared to published works in the field, but it was probably not large enough to take full advantage of CNN models. The models used in this research employed a small number of filters with a shallow architecture, and this decreased the deep learning "black box" depth and not fully fit the training data; and it reduced the computational burden, which is advantageous. A wide range of choices were available to increase the depth of the CNN architecture and optimize the training parameters. Hyperparameter optimization of the CNN architecture and training parameters were not performed because the focus here was to investigate the superiority of multimodal ensemble learning with simple models. Tuning the hyperparameters using a grid or random search method, for example, is computationally intense. A number of different optimization algorithms have been proposed (Yu and Zhu, 2020); developing an efficient scheme to optimize the hyperparameters is a topic for future investigation.

For the Brainnet CNN and MsRNN, there are unexplored options for selecting an atlas. In this study, a commonly used functional atlas was considered with few filters similar to the multimodal CNN. Correlation does not account for higher-order interactions because it is a first-order transformation (El Gazzar et al., 2019); therefore, different voxel measurements for rs-fMRI, such as entropy (Menon and Krishnamurthy, 2019b) and other connectivity measures can be investigated. The dynamic nature of the functional connectivity was not analyzed due to the increased computational requirements. Also, no comparison was performed with linear models because a voxel-wise analysis of linear models would suffer from the issues of high dimensionality.

Two strategies that may deserve attention are transfer learning and data augmentation (Vieira et al., 2017; Zhang et al., 2020). Transfer learning involves applying features learned from one dataset to tune another similar dataset. Gong et al. (2021) successfully applied transfer learning strategy exploiting big data from UK Biobank (Miller et al., 2016) in the Predictive Analysis Challenge 2019 dataset, achieving first place. Data augmentation is a strategy used in computer vision applications to enlarge the sample size by applying transformations to the data. Data augmentation methods are only now being addressed for medical imaging classification tasks, but further studies are needed for investigating disorders using 3D brain images with voxel-level data (Zhang et al., 2020).

6. CONCLUSION

The recent availability of public neuroimaging data, such as the ABCD Study, UK Biobank (Miller et al., 2016), and Child Mind Institute-Healthy Brain Network (Alexander et al., 2017), help researchers to develop novel machine learning techniques for studying brain diseases and disorders. The ensemble method with multiple modalities is ideally suited to model heterogeneity that is typical with brain abnormalities. 3D CNN together with visualization using grad-CAM is a promising way to identify neuroimaging phenotypes for the diagnosis of DBDs. Future studies are needed to investigate the use of other neuroimaging modalities to better understand the pathophysiology of brain disorders.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Adolescent Brain Cognitive Development Study (https://dx.doi.org/10.15154/1504041). Custom in-house MATLAB and Python scripts are made publicly available in GitHub (https://github.com/sreevalsansmenon/Multimodal_Ensemble).

AUTHOR CONTRIBUTIONS

SM and KK: conceptualization, methodology, formal analysis, writing–original draft preparation, and writing–review and editing. SM: software. Both authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

Data used in the preparation of this article were obtained from the Adolescent Brain Cognitive Development (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 108-120 months and follow them over 120 months into early adulthood. The ABCD Study was supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can

be found at https://abcdstudy.org/scientists/workgroups/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in analysis or writing of this report. The ABCD data used in this report came from https://dx.doi.org/10.15154/1504041.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2021.742807/full#supplementary-material

REFERENCES

Alegria, A. A., Radua, J., and Rubia, K. (2016). Meta-analysis of fMRI studies of disruptive behavior disorders. *Am. J. Psychiatry* 173, 1119–1130. doi: 10.1176/appi.ajp.2016.15081089

Alexander, M., Lindsay, E. J., Ai, L., Andreotti, C., Febre, K., Mangone, A., et al. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Sci. Data* 4:170181. doi: 10.1038/sdata.2017.181

Allen, J. L., Hwang, S., and Huijding, J. (2020). "Disruptive behavior disorders," in *The Encyclopedia of Child and Adolescent Development*, eds S. Hupp and J. Jewell (John Wiley and Sons Inc.), 1–13. doi: 10.1002/9781119171492.wecad448

American Psychiatric Association (2013). *Diagnostic and Statistical Manual of Mental Disorders, 5th Edn*. Arlington, VA: American Psychiatric Publishing.

Andersson, J. L., Jenkinson, M., and Smith, S. (2007a). *Non-Linear Optimisation*. FMRIB technical report TR07JA1, FMRIB Centre, Oxford, UK.

Andersson, J. L., Jenkinson, M., and Smith, S. (2007b). *Non-Linear Registration*. AKA Spatial Normalisation FMRIB Technical Report TR07JA2. FMRIB Analysis Group of the University of Oxford.

Barch, D. M., Albaugh, M. D., Avenevoli, S., Chang, L., Clark, D. B., Glantz, M. D., et al. (2018). Demographic, physical and mental health assessments in the adolescent brain and cognitive development study: rationale and description. *Dev. Cogn. Neurosci.* 32, 55–66. doi: 10.1016/j.dcn.2017.10.010

Cao, W., Li, C., Zhang, J., Dong, D., Sun, X., Yao, S., et al. (2019). Regional homogeneity abnormalities in early-onset and adolescent-onset conduct disorder in boys: a resting-state fMRI study. *Front. Hum. Neurosci.* 13:26. doi: 10.3389/fnhum.2019.00026

Cao, W., Sun, X., Dong, D., Yao, S., and Huang, B. (2018). Sex differences in spontaneous brain activity in adolescents with conduct disorder. *Front. Psychol.* 9:1598. doi: 10.3389/fpsyg.2018.01598

Casey, B., Cannonier, T., Conley, M. I., Cohen, A. O., Barch, D. M., Heitzeg, M. M., et al. (2018). The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. *Dev. Cogn. Neurosci.* 32, 43–54. doi: 10.1016/j.dcn.2018.03.001

Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). "Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks," in *IEEE Winter Conference on Applications of Computer Vision (WACV)* (Lake Tahoe, Nevada), 839–847. doi: 10.1109/WACV.2018.00097

El Gazzar, A., Cerliani, L., van Wingen, G., and Thomas, R. M. (2019). "Simple 1-d convolutional networks for resting-state fMRI based classification in autism," in *2019 International Joint Conference on Neural Networks (IJCNN)* (Budapest), 1–6. doi: 10.1109/IJCNN.2019.8852002

Fairchild, G., Hagan, C. C., Walsh, N. D., Passamonti, L., Calder, A. J., and Goodyer, I. M. (2013). Brain structure abnormalities in adolescent girls with conduct disorder. *J. Child Psychol. Psychiatry All. Discipl.* 54, 86–95. doi: 10.1111/j.1469-7610.2012.02617.x

Fairchild, G., Passamonti, L., Hurford, G., Hagan, C. C., Von Dem Hagen, E. A., Van Goozen, S. H., et al. (2011). Brain structure abnormalities in early-onset and adolescent-onset conduct disorder. *Am. J. Psychiatry* 168, 624–633. doi: 10.1176/appi.ajp.2010.10081184

Fairchild, G., Toschi, N., Hagan, C. C., Goodyer, I. M., Calder, A. J., and Passamonti, L. (2015). Cortical thickness, surface area, and folding alterations in male youths with conduct disorder and varying levels of callous-unemotional traits. *NeuroImage* 8, 253–260. doi: 10.1016/j.nicl.2015.04.018

Garavan, H., Bartsch, H., Conway, K., Decastro, A., Goldstein, R., Heeringa, S., et al. (2018). Recruiting the abcd sample: Design considerations and procedures. *Dev. Cogn. Neurosci.* 32, 16–22. doi: 10.1016/j.dcn.2018.04.004

Gong, W., Beckmann, C. F., Vedaldi, A., Smith, S. M., and Peng, H. (2021). Optimising a simple fully convolutional network for accurate brain age prediction in the pac 2019 challenge. *Front. Psychiatry* 12:658. doi: 10.3389/fpsyt.2021.627996

Graziano, P. A., Garic, D., and Dick, A. S. (2021). Individual differences in white matter of the uncinate fasciculus and inferior fronto-occipital fasciculus: possible early biomarkers for callous-unemotional behaviors in young children with disruptive behavior problems. *J. Child Psychol. Psychiatry All. Discipl.* doi: 10.1111/jcpp.13444. [Epub ahead of print].

Haney-Caron, E., Caprihan, A., and Stevens, M. C. (2014). DTI-measured white matter abnormalities in adolescents with conduct disorder. *J. Psychiatr. Res.* 48, 111–120. doi: 10.1016/j.jpsychires.2013.09.015

Hawes, S. W., Waller, R., Byrd, A. L., Bjork, J. M., Dick, A. S., Sutherland, M. T., et al. (2020). Reward processing in children with disruptive behavior disorders and callous-unemotional traits in the abcd study. *Am. J. Psychiatry* 178, 333–342. doi: 10.1176/appi.ajp.2020.19101092

Hubel, D. H., and Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154.2. doi: 10.1113/jphysiol.1962.sp006837

Huebner, T., Vloet, T. D., Marx, I., Konrad, K., Fink, G. R., Herpertz, S. C., et al. (2008). Morphometric brain abnormalities in boys with conduct disorder. *J. Am. Acad. Child Adolesc. Psychiatry* 47, 540–547. doi: 10.1097/CHI.0b013e3181676545

Hummer, T. A., Wang, Y., Kronenberger, W. G., Dunn, D. W., and Mathews, V. P. (2015). The relationship of brain structure to age and executive functioning in adolescent disruptive behavior disorder. *Psychiatry Res.* 231, 210–217. doi: 10.1016/j.pscychresns.2014.11.009

Hyatt, C. J., Haney-Caron, E., and Stevens, M. C. (2012). Cortical thickness and folding deficits in conduct-disordered adolescents. *Biol. Psychiatry* 72, 207–214. doi: 10.1016/j.biopsych.2011.11.017

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* 17, 825–841. doi: 10.1006/nimg.2002.1132

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156. doi: 10.1016/S1361-8415(01)00036-6

Jiang, Y., Guo, X., Zhang, J., Gao, J., Wang, X., Situ, W., et al. (2015). Abnormalities of cortical structures in adolescent-onset conduct disorder. *Psychol. Med.* 45, 3467–3479. doi: 10.1017/S0033291715001361

Kawahara, J., Brown, C. J., Miller, S. P., Booth, B. G., Chau, V., Grunau, R. E., et al. (2017). Brainnetcnn: convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage* 146, 1038–1049. doi: 10.1016/j.neuroimage.2016.09.046

Khosla, M., Jamison, K., Kuceyeski, A., and Sabuncu, M. R. (2019). Ensemble learning with 3d convolutional neural networks for functional connectome-based prediction. *NeuroImage* 199, 651–662. doi: 10.1016/j.neuroimage.2019.06.012

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (Lake Tahoe, Nevada), 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lindner, P., Savic, I., Sitnikov, R., Budhiraja, M., Liu, Y., Jokinen, J., et al. (2016). Conduct disorder in females is associated with reduced corpus callosum structural integrity independent of comorbid disorders and exposure to maltreatment. *Transl. Psychiatry* 6:e714. doi: 10.1038/tp.2015.216

Lu, F., Wang, M., Xu, S., Chen, H., Yuan, Z., Luo, L., et al. (2020). Decreased interhemispheric resting-state functional connectivity in male adolescents with conduct disorder. *Brain Imaging Behav.* 15, 1201–1210. doi: 10.1007/s11682-020-00320-8

Lu, F.-M., Zhou, J.-S., Zhang, J., Xiang, Y.-T., Zhang, J., Liu, Q., et al. (2015). Functional connectivity estimated from resting-state fmri reveals selective alterations in male adolescents with pure conduct disorder. *PLoS ONE* 10:e145668. doi: 10.1371/journal.pone.0145668

Lu, H., Liu, S., Wei, H., and Tu, J. (2020). Multi-kernel fuzzy clustering based on auto-encoder for fMRI functional network. *Expert Syst. Appl.* 159:113513. doi: 10.1016/j.eswa.2020.113513

Lundervold, A. S., and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitsch. Med. Phys.* 29, 102–127. doi: 10.1016/j.zemedi.2018.11.002

McLaughlin, K. A., Weissman, D., and Bitrán, D. (2019). Childhood adversity and neural development: a systematic review. *Annu. Rev. Dev. Psychol.* 1, 277–312. doi: 10.1146/annurev-devpsych-121318-084950

Menon, S. S., and Krishnamurthy, K. (2019a). A comparison of static and dynamic functional connectivities for identifying subjects and biological sex using intrinsic individual brain connectivity. *Sci. Rep.* 9:5729. doi: 10.1038/s41598-019-42090-4

Menon, S. S., and Krishnamurthy, K. (2019b). A study of brain neuronal and functional complexities estimated using multiscale entropy in healthy young adults. *Entropy* 21:995. doi: 10.3390/e21100995

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* 19, 1523–1536. doi: 10.1038/nn.4393

Nock, M. K., Kazdin, A. E., Hiripi, E., and Kessler, R. C. (2006). Prevalence, subtypes, and correlates of dsm-iv conduct disorder in the national comorbidity survey replication. *Psychol. Med.* 36, 699–710. doi: 10.1017/S0033291706007082

Opelt, A., Fussenegger, M., Pinz, A., and Auer, P. (2004). "Weak hypotheses and boosting for generic object detection and recognition," in *Computer Vision - ECCV 2004* (Berlin; Heidelberg: Springer), 71–84. doi: 10.1007/978-3-540-24671-8_6

O'Shea, T., and Hoydis, J. (2017). An introduction to deep learning for the physical layer. *IEEE Trans. Cogn. Commun. Netw.* 3, 563–575. doi: 10.1109/TCCN.2017.2758370

Pulini, A., Kerr, W. T., Loo, S. K., and Lenartowicz, A. (2019). Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. *Biol. Psychiatry* 4, 108–120. doi: 10.1016/j.bpsc.2018.06.003

Puzzo, I., Seunarine, K., Sully, K., Darekar, A., Clark, C., Sonuga-Barke, E. J., et al. (2018). Altered white-matter microstructure in conduct disorder is specifically associated with elevated callous-unemotional traits. *J. Abnorm. Child Psychol.* 46, 1451–1466. doi: 10.1007/s10802-017-0375-5

Rivenbark, J. G., Odgers, C. L., Caspi, A., Harrington, H., Hogan, S., Houts, R. M., et al. (2018). The high societal costs of childhood conduct problems: evidence from administrative records up to age 38 in a longitudinal birth cohort. *J. Child Psychol. Psychiatry* 59, 703–710. doi: 10.1111/jcpp.12850

Rubia, K., Smith, A. B., Halari, R., Matsukura, F., Mohammad, M., Taylor, E., et al. (2009). Disorder-specific dissociation of orbitofrontal dysfunction in boys with pure conduct disorder during reward and ventrolateral prefrontal dysfunction in boys with pure adhd during sustained attention. *Am. J. Psychiatry* 166, 83–94. doi: 10.1176/appi.ajp.2008.08020212

SAMHSA (2011). *Interventions for Disruptive Behavior Disorders: How to Use the Evidence-Based Practices Kits.* Rockville, MD: Center for Mental Health Services; Substance Abuse and Mental Health Services Administration; U.S. Department of Health and Human Services.

Sarkar, S., Dell'Acqua, F., Walsh, S. F., Blackwood, N., Scott, S., Craig, M. C., et al. (2016). A whole-brain investigation of white matter microstructure in adolescents with conduct disorder. *PLoS ONE* 11:e155475. doi: 10.1371/journal.pone.0155475

Sawilowsky, S. S. (2009). New effect size rules of thumb. *J. Modern Appl. Stat. Methods* 8, 597–599. doi: 10.22237/jmasm/1257035100

Scarmeas, N., Brandt, J., Blacker, D., Albert, M., Hadjigeorgiou, G., Dubois, B., et al. (2007). Disruptive behavior as a predictor in Alzheimer disease. *Arch. Neurol.* 64, 1755–1761. doi: 10.1001/archneur.64.12.1755

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice), 618–626. doi: 10.1109/ICCV.2017.74

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., et al. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U.S.A.* 106, 13040–13045. doi: 10.1073/pnas.0905267106

Thomas, R. M., Gallo, S., Cerliani, L., Zhutovsky, P., El-Gazzar, A., and van Wingen, G. (2020). Classifying autism spectrum disorder using the temporal statistics of resting-state functional mri data with 3D convolutional neural networks. *Front. Psychiatry* 11:440. doi: 10.3389/fpsyt.2020.00440

Vieira, S., Pinaya, W. H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* 74, 58–75. doi: 10.1016/j.neubiorev.2017.01.002

Wallace, G. L., White, S. F., Robustelli, B., Sinclair, S., Hwang, S., Martin, A., et al. (2014). Cortical and subcortical abnormalities in youths with conduct disorder and elevated callous-unemotional traits. *J. Am. Acad. Child Adolesc. Psychiatry* 53, 456–465.e1. doi: 10.1016/j.jaac.2013.12.008

Waller, R., Hawes, S. W., Byrd, A. L., Dick, A. S., Sutherland, M. T., Riedel, M. C., et al. (2020). Disruptive behavior problems, callous-unemotional traits, and regional gray matter volume in the adolescent brain and cognitive development study. *Biol. Psychiatry* 5, 481–489. doi: 10.1016/j.bpsc.2020.01.002

Werhahn, J. E., Mohl, S., Willinger, D., Smigielski, L., Roth, A., Hofstetter, C., et al. (2020). Aggression subtypes relate to distinct resting state functional connectivity in children and adolescents with disruptive behavior. *Eur. Child Adolesc. Psychiatry* 30, 1237–1249. doi: 10.1007/s00787-020-01601-9

Wu, Q., Zhang, X., Dong, D., Wang, X., and Yao, S. (2017). Altered spontaneous brain activity in adolescent boys with pure conduct disorder revealed by regional homogeneity analysis. *Eur. Child Adolesc. Psychiatry* 26, 827–837. doi: 10.1007/s00787-017-0953-7

Yan, W., Calhoun, V., Song, M., Cui, Y., Yan, H., Liu, S., et al. (2019). Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site fMRI data. *EBioMedicine* 47, 543–552. doi: 10.1016/j.ebiom.2019.08.023

Yu, T., and Zhu, H. (2020). Hyper-parameter optimization: a review of algorithms and applications. *arXiv [Preprint]. arXiv:2003.05689.* Available online at: https://arxiv.org/abs/2003.05689

Zhang, J., Li, B., Gao, J., Shi, H., Wang, X., Jiang, Y., et al. (2015). Impaired frontal-basal ganglia connectivity in male adolescents with conduct disorder. *PLoS ONE* 10:e145011. doi: 10.1371/journal.pone.0145011

Zhang, J., Zhou, J., Lu, F., Chen, L., Huang, Y., Chen, H., et al. (2017). Investigation of the changes in the power distribution in resting-state brain networks associated with pure conduct disorder. *Sci. Rep.* 7:5528. doi: 10.1038/s41598-017-05863-3

Zhang, L., Wang, M., Liu, M., and Zhang, D. (2020). A survey on deep learning for neuroimaging-based brain disorder analysis. *Front. Neurosci.* 14:779. doi: 10.3389/fnins.2020.00779

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., and Slaney, G. (2021). Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance

imaging. *J. Neurosci. Methods* 353:109098. doi: 10.1016/j.jneumeth.2021.109098

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, Nevada).

Author Disclaimer: This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 10 December 2021
doi: 10.3389/fninf.2021.670052

Analyzing Complex Problem Solving by Dynamic Brain Networks

Abdullah Alchihabi[1], Omer Ekmekci[1], Baran B. Kivilcim[1], Sharlene D. Newman[2] and Fatos T. Yarman Vural[1]*

[1] Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, [2] Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN, United States

Complex problem solving is a high level cognitive task of the human brain, which has been studied over the last decade. Tower of London (TOL) is a game that has been widely used to study complex problem solving. In this paper, we aim to explore the underlying cognitive network structure among anatomical regions of complex problem solving and its subtasks, namely *planning* and *execution*. A new computational model for estimating a brain network at each time instant of fMRI recordings is proposed. The suggested method models the brain network as an Artificial Neural Network, where the weights correspond to the relationships among the brain anatomic regions. The first step of the model is preprocessing that manages to decrease the spatial redundancy while increasing the temporal resolution of the fMRI recordings. Then, dynamic brain networks are estimated using the preprocessed fMRI signal to train the Artificial Neural Network. The properties of the estimated brain networks are studied in order to identify regions of interest, such as hubs and subgroups of densely connected brain regions. The representation power of the suggested brain network is shown by decoding the planning and execution subtasks of complex problem solving. Our findings are consistent with the previous results of experimental psychology. Furthermore, it is observed that there are more hubs during the planning phase compared to the execution phase, and the clusters are more strongly connected during planning compared to execution.

Keywords: fMRI, machine learning, brain networks, tower of London (TOL), complex problem solving

OPEN ACCESS

Edited by:
Xiang Li,
Massachusetts General Hospital and Harvard Medical School, United States

Reviewed by:
Haixing Dai,
University of Georgia, United States
Bao Ge,
Shaanxi Normal University, China

****Correspondence:***
Abdullah Alchihabi
abdullahalchihabi@cmail.carleton.ca

Received: *20 February 2021*
Accepted: *10 November 2021*
Published: *10 December 2021*

Citation:
Alchihabi A, Ekmekci O, Kivilcim BB, Newman SD and Yarman Vural FT (2021) Analyzing Complex Problem Solving by Dynamic Brain Networks. Front. Neuroinform. 15:670052. doi: 10.3389/fninf.2021.670052

1. INTRODUCTION

Complex problem solving is a very crucial ability of the human brain, which covers a large number of high-level cognitive processes, including strategy formation, coordination, sequencing of mental functions, and holding information online. These complex high-level cognitive processes make the inner workings of problem solving a challenging task.

The standard method for neuro-analysis of complex problem solving in the literature is to study the fMRI data recorded while the subjects play the Tower of London (TOL) game, designed by Shallice (1982). TOL game consists of three bins having different capacities with colored balls placed in the bins; the aim is to rearrange the balls from their initial state to a predetermined goal state while moving one ball at a time and taking into consideration the limited capacity of each bin (as shown in **Figure 1**).

TOL game has been primarily employed to study the effect of various properties of complex problem solving performance in healthy subjects. The predictive power of working memory, inhibition, and fluid intelligence on TOL performance has been investigated with consideration

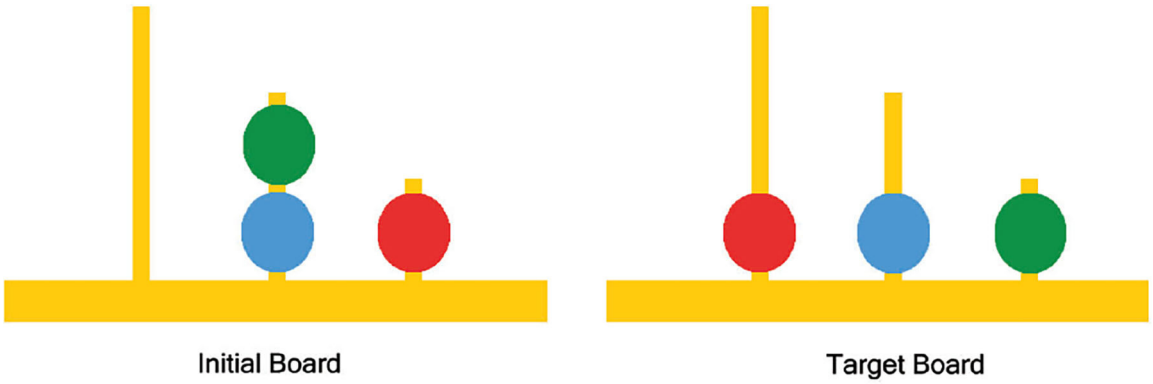


FIGURE 1 | Example tower of London (TOL) puzzle.

of factors such as age, gender, exercise, etc. (Unterrainer et al., 2004, 2005; Zook et al., 2004, 2006; Boghi et al., 2006; Albert and Steinberg, 2011; Chang et al., 2011; Desco et al., 2011; Kaller et al., 2012). Additionally, TOL has been used to investigate the effect of various clinical disorders on functions associated with the prefrontal cortex such as planning. For example, the task has been utilized in neuroimaging studies to identify executive dysfunction by examining differential cognitive activation patterns in people suffering from neurological disorders like epilepsy, seizures, depression, Parkinson's and schizophrenia (Goethals et al., 2005; Rasser et al., 2005; Rektorova et al., 2008; MacAllister et al., 2012).

The classic work of Newell and Simon (Newell et al., 1957; Simon and Newell, 1971) hypothesized three distinct phases of complex problem solving: construction of problem representation, elaboration to search for operators to solve the problem, and execution to implement the solution. Despite being a well respected theory, there is little to no evidence from cognitive neuroimaging that supports this hypothesis directly. Consequently, refinements of the theory, such as online planning in which elaboration and execution phases are interspersed, and mechanisms such as schema development that may suggest qualitative differences between good and poor problem-solvers, are understood even less. The primary reason for this state of affairs is that cognitive neuroimaging in general and fMRI analysis, in particular, tends to ignore the temporal aspect of how brain activation and network connectivity evolve during complex cognitive tasks. Further, much of the existing methods have tended to test theoretical cognitive models by searching for brain data that fit those models, rather than using the brain data themselves to inform us about cognition.

Numerous studies have proposed various computational models in order to build brain networks from fMRI measurements, both during cognitive tasks or during resting state. These studies represent a shift in the literature toward brain decoding algorithms that are based on the connectivity patterns in the brain motivated by the findings that these patterns provide more information about cognitive tasks than the isolated behavior of individual groups of voxels or anatomical regions (Lindquist, 2008; Ekman et al., 2012; Shirer et al., 2012; Richiardi et al., 2013; Onal et al., 2017). Some of these studies focused on the pairwise relationships between voxels or brain regions. For example, Pearson correlation has been used in order to construct undirected functional connectivity graphs at different frequency resolutions in Richiardi et al. (2011). Also, pairwise correlations and mutual information have been used in order to build functional brain networks in various studies aiming to investigate the network differences between patients with Schizophrenia or Alzheimer's disease and healthy subjects (Lynall et al., 2010; Menon, 2011; Kurmukov et al., 2017). Others used partial correlation along with constrained linear regression to generate brain networks in Lee et al. (2011).

In our previous studies, we take advantage of the locality property of the brain by constructing local mesh networks around each brain region. Then, we represent the entire brain network as an ensemble of local meshes. In these studies, we estimated the Blood-Oxygenation Level Dependent (BOLD) response of each brain region as a linear combination of the responses of its "closest" neighboring regions. Then, we solved the systems of linear equations using various regression techniques. Our team applied Levinson-Durbin recursion in order to estimate the edge weights of each local star mesh, where the nodes are the neighboring regions of the seed brain region (Fırat et al., 2013; Alchihabi et al., 2018). We also used ridge regression to estimate edge weights while constructing the local mesh networks across windows of time series of fMRI recordings (Onal et al., 2015, 2017).

In this study, we present a novel approach for estimating dynamic brain networks, which represent the relationship among the brain anatomic regions at each time instant of the fMRI recordings. The approach models the relationship among the anatomical brain regions as an Artificial Neural Network (ANN), where the edge weights correspond to the arc weights of the brain network. The idea of modeling the brain network as an ANN is first introduced in our lab (Kivilcim et al., 2018), where the model can be constructed to estimate both directed and undirected brain graphs. In this study, we further extend this idea to estimate dynamic brain networks. We also explore the

validity and representation power of the suggested brain network by analyzing its statistical properties using the methods suggested in Bassett and Bullmore (2006), Power et al. (2010), Rubinov and Sporns (2010), and Park and Friston (2013).

Several network measures, such as measures of centrality, which identify potential hubs and measures of functional segregation, which detect densely interconnected clusters of nodes, provide means to analyze both individual components of brain network and the brain network as a whole. As a result, these network measures reveal and characterize various aspects of inter-dynamics of brain regions enabling us to analyze and compare different brain network snapshots. The properties of the dynamic brain networks are studied in order to identify the active anatomical regions during both planning and execution phases of complex problem solving. Potential hubs and clusters of densely connected brain regions are identified for both subtasks. Furthermore, the distinctions and similarities between planning and execution networks are highlighted. The results identify both active and inactive hub regions as well as clusters of densely connected anatomical regions during complex problem-solving. In addition, results show that there are more potential hubs during the planning phase compared to the execution phase. Also, the clusters of densely interconnected regions are significantly more strongly connected during planning compared to execution. Finally, we studied the decoding power of the suggested brain network model by using simple machine learning methods to classify two phases of complex problem solving, namely, planning and execution.

2. TOL EXPERIMENT PROCEDURE

In this section, we introduce the details of the experiment as well as data collection and preprocessing methods.

2.1. Participants and Stimuli

18 college students aged between 19 and 38 participated in the experiment after signing informed, written consent documents approved by the Indiana University Institutional Review Board. The subjects solved a computerized version of TOL problem; two configurations were presented at the beginning of each puzzle: the initial state and the goal state. The subjects were asked to transform the initial state into the goal state using the minimum number of moves. However, the subjects were not informed of the minimum number of moves needed to solve a given puzzle nor of the existence of multiple solution paths.

2.2. Procedure

Each subject underwent a practice session before entering the scanning session to acquaint subjects with the TOL problem. The subjects were given the following instructions: "You will be asked to solve a series of puzzles. The goal of the puzzle is to make the 'start' or 'current' state match the 'goal' state (They were shown an example). Try to solve the problems in the minimum number of moves by planning ahead. Work as quickly and accurately as possible, but accuracy is more important than speed."

The scanning session consisted of 4 runs, each run included 18 timed puzzles, with a 5-s planning only time slot during which subjects were not allowed to move the balls. However, they were allowed to continue planning after the 5 s planning only time slot if they chose to do so. Following every puzzle, there was a 12-s rest period where subjects focused on a plus sign in the center of the screen. Each run was also followed by a 28-s fixation period.

The planning task is defined from the start of the puzzle until the subject's first move. The execution task is defined from the subject's first move until the end of the puzzle. During the experiments, both planning and execution times change across the subjects, runs and puzzles. The average planning time instances per puzzle is 5.91 and the average time instances for execution per puzzle is 5.63 over all puzzles and subjects. The average total planning time instances per run is 106.35 and the average total execution time instances per run is 101.28 over all subjects. **Figure 2** shows the total number of planning and execution instances for each run of all subjects. Each mark on the horizontal axis refers to a given run of a specific subject. As it can be observed from **Figure 2**, the data is quite balanced between the planning and execution phases (average of 101 planning and 106 execution across the subjects per each run). For this reason, we did not augment the classes or eliminate some samples in the dataset for balancing the classes. The details of the dataset are summarized in **Table 1**.

2.3. fMRI Data Acquisition and Preliminary Analysis

The fMRI images were collected using a 3 T Siemens TRIO scanner with an 8-channel radio frequency coil located in the Imaging Research Facility at Indiana University. The images were acquired in 18 5 mm thick oblique axial slices using the following set of parameters: TR = 1,000 ms, TE = 25 ms, flip angle = 60°, voxel size = 3.125 mm×3.125 mm×5 mm with a 1 mm gap.

The statistical parametric mapping toolbox was used to perform the preliminary data analysis that included: image correction for slice acquisition timing, resampling, spatial smoothing, motion correction and normalization to the Montreal Neurological Institute (MNI) EPI template. Further details concerning the procedure and data acquisition can be found in Newman et al. (2009) as we use the same data/participants in this study. It is also worth noting that we perform our analysis on all recorded puzzles, not only correctly solved ones, given that the aim of this study is to investigate the planning and execution networks in general. In future work, we aim to study the differences in the planning and execution networks between good problem-solvers and bad problem-solvers; in that case, we will make the distinction between correctly solved puzzles and unsolved puzzles. Furthermore, the entirety of our analysis is performed on the raw fMRI recordings; no first-level modeling or regressors are applied; rather, we use the recorded time series as our raw BOLD response.

In order to investigate the inter-subject variability, we estimate the mean values and variance of BOLD activation of each brain anatomic region across all subjects. **Figure 3** clearly shows the relatively low variations of the BOLD activation around the mean values of brain anatomic regions across 18 subjects.

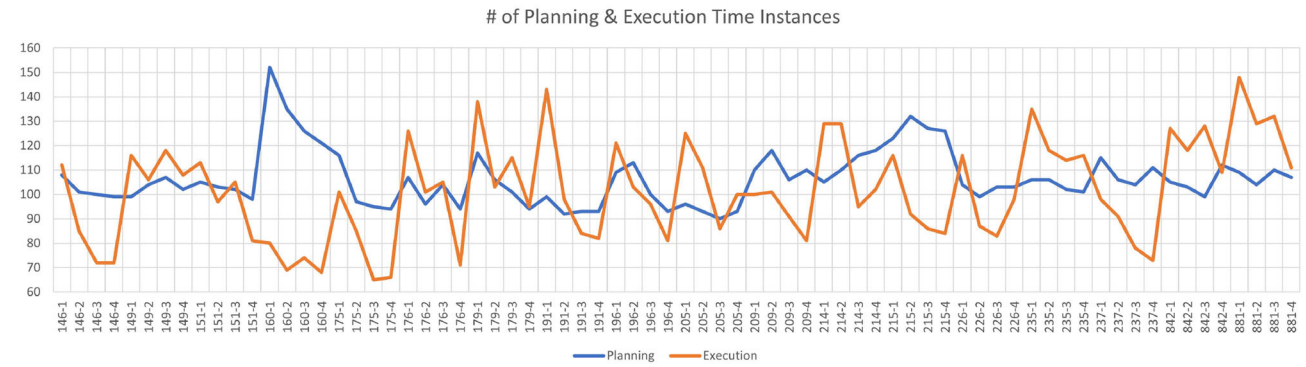


FIGURE 2 | The number of planning and execution instances for each run of all subjects. Horizontal axis indicates the subject ID for each run (there are total of 18 subject x4 runs = 72 runs), whereas the vertical axis shows the number of time instances for planning (blue) and execution (orange) per run.

TABLE 1 | Summary of TOL dataset.

# of Subjects	18
# of Runs/subject	4
# of Puzzles/run	18
Avg. Planning instances/puzzle	5.91
Avg. Execution instances/puzzle	5.63
Avg. Planning instances/run	106.35
Avg. Execution instances/run	101.28
Total planning instances for 18 x 4 runs	7,657
Total execution instances for 18 x 4 runs	7,292

Note that the planning instances (7657) and execution instances (7292) are quite balanced.

3. MODELING DYNAMIC BRAIN NETWORK AS AN ARTIFICIAL NEURAL NETWORK

Can we model the relationship among the anatomic regions as an Artificial Neural Network? If so, what is the validity and representation power of this network to analyze cognitive tasks such as complex problem solving? In this section, we suggest a computational model to represent the complex problem solving task as a dynamic brain network. In the next section, we shall explore the validity of this network and try to analyze the complex problem solving task of the human brain.

3.1. Preprocessing of the fMRI Recordings

In order to be able to estimate a dynamic brain network among the anatomic regions, we need to process the raw fMRI data for

- representation of anatomic regions,
- interpolation in time,
- injecting additive noise,

as explained below.

3.1.1. Representation of Anatomic Regions

Each anatomic region is represented by a time series using voxel selection and averaging methods. Voxel selection reduces the dimension of fMRI data (185,000 voxels per brain volume) and eliminates the irrelevant voxels that do not contribute to the underlying cognitive process. ANOVA method is used to choose the most discriminative voxels and to discard the remaining ones (Cox and Savoy, 2003; Pereira et al., 2009; Afrasiyabi et al., 2016). The f-value score of each voxel $\boldsymbol{v_i}$ is calculated from Equation (1):

$$f\_score_i = \frac{MSB(\boldsymbol{v_i}, \boldsymbol{y_{label}})}{MSW(\boldsymbol{v_i}, \boldsymbol{y_{label}})}, \tag{1}$$

where $\boldsymbol{y_{label}}$ is the label indicating the subtask (Planning or Execution). $MSB(\boldsymbol{v_i}, \boldsymbol{y_{label}})$ is the mean square value between raw measured BOLD response of voxel i and the label vector $\boldsymbol{y_{label}}$, which is calculated by Equation (2):

$$MSB(\boldsymbol{v_i}, \boldsymbol{y_{label}}) = \frac{SSB(\boldsymbol{v_i}, \boldsymbol{y_{label}})}{df_{between}}, \tag{2}$$

$SSB(\boldsymbol{v_i}, \boldsymbol{y_{label}})$ is the sum of squares between $\boldsymbol{y_{label}}$ and $\boldsymbol{v_i}$, $df_{between}$ is the number of groups minus one. $MSW(\boldsymbol{v_i}, \boldsymbol{y_{label}})$ is the mean square value within voxel i and the label vector $\boldsymbol{y_{label}}$ and it is calculated by Equation (3):

$$MSW(\boldsymbol{v_i}, \boldsymbol{y_{label}}) = \frac{SSW(\boldsymbol{v_i}, \boldsymbol{y_{label}})}{df_{within}}, \tag{3}$$

where $SSW(\boldsymbol{v_i}, \boldsymbol{y_{label}})$ is the sum of squares within group and df_{within} is the degree of freedom within (total number of elements in $\boldsymbol{v_i}$ and $\boldsymbol{y_{label}}$ minus the number of groups).

We order the voxels according to their f-value scores. Then, the distribution of f-value scores of all voxels is plotted in order to determine the number of voxels to retain. Voxel selection is applied to the voxels of all brain regions except the ones located in the cerebellum, which we exclude during network extraction.

Each anatomic region is represented by averaging the BOLD response of the selected voxels, which resides in that region defined by automated anatomical labeling (AAL) atlas (Tzourio-Mazoyer et al., 2002) as shown in Equation (4):

$$\boldsymbol{r_j} = \frac{\sum_{i \in \zeta[j]} \boldsymbol{v_i}}{|\zeta[j]|} \tag{4}$$

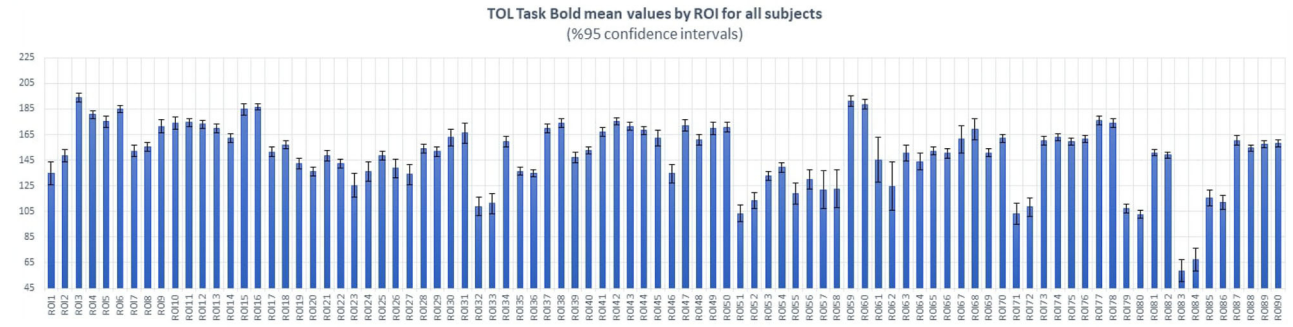


FIGURE 3 | Estimated mean values and variations of BOLD activation of each brain anatomic region, over 18 subjects. Horizontal axis shows the index of anatomic regions, whereas the vertical axis shows the mean values (blue bars) and variances (black bars) of measured BOLD activation.

where $\boldsymbol{r_j}$ is the representative BOLD response of region j, $\boldsymbol{v_i}$ is the raw measured BOLD response of voxel i and $\zeta[j]$ is the set of selected voxels located in region j. The representative BOLD responses,$\boldsymbol{r_j}$, enables us to investigate the role and contribution of each region to the planning and execution phases of the problem solving task.

3.1.2. Interpolation

It is well-known that despite its high spatial resolution, fMRI signal has very low temporal resolution compared to EEG signal. In this study, we interpolate the fMRI signal in order to compensate for this drawback and study the effect of interpolation for estimating the brain networks and on decoding the planning and execution phases of TOL game.

In the TOL study, subjects solved a puzzle in at most 15 s and the sampling rate, TR, is 1,000 ms. Interpolation is used to increase the temporal resolution by estimating z extra brain volumes between each two consecutive measured brain volumes. As a result, the total number of available brain volumes for each puzzle becomes $n + z * (n - 1)$, where n is the number of measured brain volumes of a given puzzle. We use the cubic spline interpolation function rather than linear interpolation methods in order to prevent edge effects and smoothing out the spikes between the measured brain volumes (McKinley and Levine, 1998).

In order to analyze the effect of time interpolation and to estimate an acceptable number of inserted brain volumes z, we compare the Fourier Transform of the fMRI signal computed before and after interpolation so that the frequency content of the signal is not distorted by interpolation. The original single-sided amplitude of the signal and the one obtained after interpolation are compared in order to ensure that interpolation is preserving the smooth peaks of the data in the frequency domain (Cochran et al., 1967; Frigo and Johnson, 1998).

3.1.3. Injecting Gaussian Noise

When modeling a deterministic signal by a probabilistic method, adding noise to the signal decreases the estimation error in most of the practical applications. The final phase of preprocessing is adding Gaussian noise to the interpolated time series of the BOLD response in each anatomical region. For this purpose, instead of just injecting white noise, a rather informed noise, *colorful Gaussian noise*, is added. In order to reflect the corresponding brain region's properties, for each sample, the additive noise sample is generated from a Gaussian distribution having mean and variance of that anatomical region. These newly generated samples not only act like a natural regularizer to improve the generalization performance of brain decoding but also help making the Artificial Neural Network more stable when estimating the edge weights of the brain networks (Matsuoka, 1992; Reed et al., 1992).

Given a representative time series from a particular brain region, i represents the index of an anatomical region. The new samples are generated with vector addition of noise while preserving the signal-to-noise ratio (SNR) as in $\tilde{\boldsymbol{r}}_{\boldsymbol{j}} = \boldsymbol{r_j} + \boldsymbol{\tau_j}$, where $\boldsymbol{\tau_j}$ is a noise vector sampled from $N(\alpha_{noise}\ \mu(\boldsymbol{r_j}),\ \beta_{noise}\ \sigma^2(\boldsymbol{r_j}))$, α_{noise} and β_{noise} are the scaling factors which are set empirically, to optimize the decoding performance.

3.2. Building Dynamic Brain Networks With Artificial Neural Networks

The above preprocessing methods yield a relatively high temporal resolution and smooth time series for each anatomic region compared to the row fMRI recordings.

In this section, we use the output of the preseprocessing step to estimate the relationship among the time series of anatomic regions at each time instance to generate a dynamic brain network, where the arc weights vary with respect to time instances.

3.2.1. Partitioning the Time Series Into Fixed Size Internals and Defining the Brain Network for Each Window

As the first step, we partition each time series, which represents an anatomical region, into fixed-size windows. Each window, $win(t)$, is centered at the measured brain volume at time instance, t. The size of each window is $Win\_Size = z + 1$ brain volumes, where z is the number of interpolated brain volumes in each window. Equation (5) shows the time instances included in each window.

$$\boldsymbol{win(t)} = \left[t - \left\lfloor \frac{z}{2} \right\rfloor, \ldots, t, \ldots, t + \left\lceil \frac{z}{2} \right\rceil \right] \quad (5)$$

We define a dynamic brain network, $N(t) = (V, W(t))$, for each time window $win(t)$, where V is the set of nodes of the graph corresponding to the brain anatomical regions and $W(t) = \{w_{t,j,i} | \forall i, j \in V\}$ is the directed weighted edges between the nodes of the graph within time window $win(t)$. The nodes of the graph represent the AAL-defined brain regions (Tzourio-Mazoyer et al., 2002), except for the regions located in the cerebellum. The nodes are then pruned using voxel selection, as some anatomical regions contribute no voxels at all and get deleted from the set of nodes of the graph V.

Note that our aim is to label the BOLD responses measured at each brain volume as it belongs to one of the two phases of complex problem solving, namely, *planning and execution*. For this purpose, we represent each brain volume measured at a time instant t by a network, which shows the relationship among the anatomical regions. This dynamic network representation will allow us to investigate the network properties of planning and execution subtasks.

Note also that the nodes, V, of the network are fixed to the active anatomic regions, and our goal is only to estimate the weights of the edges, $W(t)$, of the brain network, $N(t)$, for each time instance, t For this purpose we adopt the method suggested by our team in Kivilcim et al. (2018).

3.2.2. Forming Local Meshes

It is well-known that the human brain operates with two contradicting principles, namely *locality* and *centrality*. Our suggested network model incorporates these two principals by defining a set of spatially local meshes then ensembling the local meshes to form the brain network. This representation not only avoids to define fully connected brain networks by omitting the connectivity among irrelevant brain regions but also reduces the computational complexity.

In order to define local meshes, for each window $\boldsymbol{win(t)}$, we define the functional neighborhood matrix, Ω_t, for each time instant t. The entries of Ω_t are binary, either 1 or 0, indicating if there is a connection between two regions or not. The size of the matrix is $M \times M$, where M is the number of brain anatomical regions. The functional neighborhood matrix contains no self-connections, thus, $\Omega_t(i, i) = 0 \, \forall i \in [1, M]$. Recall that the brain regions are pruned by voxel selection. Thus, the regions which do not contain any voxels have no in/out connections, and the corresponding entries in Ω_t are all zero.

The connectivity of each region to the rest of the regions is determined by using Pearson correlation, as follows: first, for every region i, we measure the Pearson correlation between its BOLD response $\boldsymbol{r_{i,t}}$ and the BOLD responses of all the other remaining regions as shown below:

$$cor(\boldsymbol{r_{i,t}}, \boldsymbol{r_{j,t}}) = \frac{cov(\boldsymbol{r_{i,t}}, \boldsymbol{r_{j,t}})}{\sigma(\boldsymbol{r_{i,t}})\sigma(\boldsymbol{r_{j,t}})}, \quad (6)$$

where $\boldsymbol{r_{i,t}}$ is the BOLD response of region i across time window $win(t)$, $cov(\boldsymbol{r_{i,t}}, \boldsymbol{r_{j,t}})$ is the covariance between the corresponding BOLD responses of regions i and j. σ is the standard deviation of the BOLD response of a given region. Thus, the higher the Pearson correlation between two regions the closer they are to each other in the functional neighborhood system.

Then, we select p of the regions with the highest correlation scores with region i. Thus, a local mesh for each anatomic region i is formed by obtaining the neighborhood set $\eta_p[i]$, which contains the p closest brain regions to region i. The degree of neighborhood, p, is determined empirically as will be explained in the next section. Finally, we define the $\Omega_t(i, j)$ as the connectivity between the regions i and j, using the constructed neighborhood sets as follows:

$$\Omega_t(i, j) = \begin{cases} 1, & \text{if } j \in \eta_p[i] \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Note that each anatomical region is connected to its p closest functional neighbors. This approach forms a star mesh around each anatomical region.

The ensemble of all of the local meshes creates a brain network at each time instance. Note, also, that Pearson correlation values are not used as the weights between two regions. They are just used to identify the nodes of each local mesh formed around an anatomical region. The estimated brain network becomes sparser as p gets smaller. When p is set to the number of anatomic regions, M, the network becomes fully connected. This approach of defining the connectivity matrix makes the network representation sparse for small p-values and constructs a network that is connected in functionally closest regions, satisfying the locality property of the human brain.

3.2.3. Estimating the Edge Weights of the Brain Network

After having determined the edges of the brain graph using the functional neighborhood matrix Ω_t, all that is left is to estimate the weights of these edges at each local mesh. At this point, we could use the Pearson correlation values as edge weights between two anatomic regions. However, Pearson values are restricted to measure the connectivity among the pairs only. A better approach is to consider the multiple relationships among an anatomic region and all of its neighbors in the local mesh. In order to estimate the edge weights in a mesh all at once, we represent the time series of each region i ($\boldsymbol{r_{i,t}}$) as a linear combination of its closest p-functional neighbors as shown in Equation (8):

$$\hat{\boldsymbol{r}}_{\boldsymbol{i,t}} = \sum_{j \in \eta_p[i]} w_{t,j,i} \boldsymbol{r_{j,t}} + \epsilon_{i,t}. \quad (8)$$

In Equation (8), $\hat{\boldsymbol{r}}_{\boldsymbol{i,t}}$ is the representative time series of of region i within the time window $win(t)$, $w_{t,j,i}$ is the estimated edge weight between node (region) i and node j at time instance t. $\eta_p[i]$ is the p-closest functional neighbors of region i.

Ertugrul et al. (2016) showed that representing the time series of an anatomic region as a linear combination of its closest neighbors provides better performance compared to using pairwise Pearson correlation in brain decoding. They estimated the arc-weights for each mesh formed around region i for each time window $win(t)$ by minimizing the mean-squared error loss function using Ridge regression. In this approach, the mean-squared error loss function is minimized with respect to $w_{t,j,i}$, for each mesh, independent of the other meshes, where the expectation is taken over the time-instances, in window $win(t)$ as shown in Equation (9).

$$E[(\epsilon_{i,t})^2] = E[(\hat{\boldsymbol{r}}_{\boldsymbol{i,t}} - \sum_{j \in \eta_p[i]} w_{t,j,i} \boldsymbol{r}_{\boldsymbol{j,t}})^2] + \lambda ||w_{t,j,i}||^2, \quad (9)$$

where λ is the L2 regularization parameter whose value is optimized using cross-validation. L2 regularization is used in order to improve the generalization of the constructed mesh networks. Note that the estimated arc-weights, $w_{t,j,i} \neq w_{t,i,j}$. Therefore, the ensemble of meshes yields a directed brain network.

In this study, we define an Artificial Neural Network to estimate the values of mesh arc-weights for all anatomical regions jointly in each time window, as proposed in Kivilcim et al. (2018). In this method, we estimate the mesh arc-weights matrix $W(t) = \{w_{t,j,i} | j, i \in V\}$ using a feed-forward neural network. The architecture of this network consists of an input layer and an output layer, both containing M nodes corresponding to each anatomic region. The edges of the feed-forward neural network are constructed using the neighborhood matrix Ω_t. There is an edge between node i of the output layer and node j from the input layer, if $\Omega_t(i,j) = 1$.

The loss function of the suggested Artificial Neural Network is given in Equation (10), where W is the weight matrix of the entire neural network that corresponds to directed edge weights of the brain graph and W_i is the row of matrix W corresponding to region i:

$$\begin{aligned} Loss(Output_i) &= E[(\epsilon_{i,t})^2] + \lambda W_i^T W_i \\ &= E[(\boldsymbol{r}_{\boldsymbol{i,t}} - \sum_{j \in \eta_p[i]} w_{t,j,i} \boldsymbol{r}_{\boldsymbol{j,t}})^2] + \lambda W_i^T W_i. \end{aligned} \quad (10)$$

We train the aforementioned Artificial Neural Network in order to obtain the weights of the brain network at each time instance t that minimize the loss function by applying a gradient descent optimization method as shown in Equation (11),

$$w_{t,j,i}^{(\kappa)} = w_{t,j,i}^{(\kappa-1)} - \alpha_{learning} \frac{\partial E[(\epsilon_{i,t})^2]}{\partial w_{t,j,i}}, \quad (11)$$

where $w_{t,j,i}^{(\kappa)}$ is the weight of the edge from node j to node i at epoch (iteration) κ, $\alpha_{learning}$ is the learning rate. The number of epochs and learning rate used to train the network are optimized empirically using cross-validation.

Finally, the weights of the above artificial neural network, computed for each $win(t)$, correspond to the edge weights of the dynamic brain network, $N(t) = (V, W(t))$, at each time instant t. Thus, we refer to the brain networks using their window indices in order to obtain a set of dynamic brain networks $T = \{N(1), N(2), \ldots N(tot\_win)\}$, where $N(t)$ is the brain network for time window $win(t)$ and tot_win is the total number of time windows.

3.3. Network Metrics for Analyzing Brain Networks

In this section, we introduce some measures which we will use to investigate the network properties of each phase of the complex problem solving task, namely, planning and execution, using the estimated dynamic brain functional networks. The connectivity patterns of anatomical regions are analyzed by the set of network measures given below. Two separate sets of measures are used, namely, measures of centrality and segregation. Since our estimated brain networks are directed, we distinguish the incoming and outgoing edges in the network while defining the measures.

Recall that the suggested brain network $N(t) = (V, W(t))$ consists of a set of nodes, V, each of which corresponding to one of the M anatomical regions. $W(t)$ is the dynamic edge weight matrix with the entries, $w_{i,j}$, representing the weight of the edge from node i to node j. For the sake of simplicity, we omit the time dependency parameter t, since we compute the network properties at each time instant.

3.3.1. Measures of Centrality

Measures of centrality aim to identify brain regions that play a central role in the flow of information in the brain network or nodes that can be identified as hubs. It is commonly measured using node degree, node strength and node betweenness centrality, which are defined below.

3.3.1.1. Node Degree

The degree of a node is the total number of its edges as shown in Equation (12), where $degree_i$ is the degree of node i, V is the set of all nodes in the graph and $a_{i,j}$ is the edge between node i and node j.

$$degree_i = \sum_{j \in V} a_{i,j}, \quad (12)$$

where $a_{i,j}$ takes value 0 if $(w_{i,j} == 0)$ and takes value 1 otherwise.

In the case of a directed graph, we distinguish two different metrics: node in-degree $degree_i^{in}$ and node out-degree $degree_i^{out}$ metrics which are shown in Equations (13) and (14), respectively, where $a_{j,i} = 1$, if there is a directed edge from node j to node i.

$$degree_i^{in} = \sum_{j \in V} a_{j,i} \quad (13)$$

$$degree_i^{out} = \sum_{j \in V} a_{i,j} \tag{14}$$

Node degree is a measure of centrality of the given nodes, where it aims to quantify the hub brain regions interacting with a large number of brain regions. Thus, a node with high degree indicates its central role in the network.

3.3.1.2. Node Strength

Node strength is the sum of the weights of edges connected to a given node (Equation 15), where $w_{i,j}$ is the weight of the edge between node i and node j.

$$strength_i = \sum_{j \in V} w_{i,j} \tag{15}$$

Similar to node degree, node strength, also, distinguishes two metrics in the case of directed graphs, namely, node in-strength $strength_i^{in}$ and out-strength $strength_i^{out}$ shown in Equations (16) and (17), respectively, where $w_{j,i}$ is the weight of the edge from node j to node i.

$$strength_i^{in} = \sum_{j \in V} w_{j,i} \tag{16}$$

$$strength_i^{out} = \sum_{j \in V} w_{i,j}. \tag{17}$$

Node strength is a node centrality measure that is similar to node degree, which is used in the case of weighted graphs. Nodes with large strength values are tightly connected to other nodes in the network forming hub nodes.

3.3.1.3. Node Betweenness Centrality

Betweenness centrality of node i is the fraction of the shortest paths in the network that pass through node i as shown in Equation (18)

$$betweenness_i = \frac{1}{(M-1)(M-2)} \sum_{j,k \in V} \frac{\rho_{j,k}^i}{\rho_{j,k}}, \tag{18}$$

where $\rho_{j,k}$ is the number of shortest paths between nodes j and k, $\rho_{j,k}^i$ is the number of shortest paths between nodes j and k that pass through node i, nodes i, j and k are distinct nodes.

Before measuring the betweenness centrality of a node, we need to change our perspective from connection weight matrix to connection length matrix since betweenness centrality is a distance-based metric. In connection weights matrix, larger weights imply higher correlation and shorter distance while it is the opposite in the case of length matrix. Connection length matrix is obtained by inverting the weights of the connection weight matrix. Then, the algorithm suggested in Brandes (2001) is employed in order to calculate the node betweenness centrality for each anatomical region.

Nodes with high betweenness centrality are expected to participate in many of the shortest paths of the networks. Thus, taking a crucial role in the information flow of the network.

3.3.2. Measures of Segregation

Measures of segregation aim to quantify the existence of subgroups within brain networks, where the nodes are densely interconnected. These subgroups are commonly referred to as clusters or modules. The existence of such clusters in functional brain networks is a sign of interdependence among the nodes forming the cluster. Measures of segregation include clustering coefficient, transitivity and local efficiency. While global efficiency is a measure of functional integration representing how easy it is for information to flow in the network.

3.3.2.1. Clustering Coefficient

The clustering coefficient of a node i is the fraction of triangles around node i which is calculated by Equation (19) as proposed in Fagiolo (2007). It is defined as the fraction of the neighbors of node i that are also neighbors of each other.

$$C_i = \frac{\chi_i}{[(d_i^{out} + d_i^{in})(d_i^{out} + d_i^{in} - 1) - 2\sum_{j \in V} a_{i,j}a_{j,i}]}. \tag{19}$$

where d_i^{in} is the in-degree of node i and d_i^{out} is the out-degree of node i. χ_i is the weighted geometric mean of triangles around node i that is calculated by Equation (20). Recall that $a_{j,i} = 1$, if there is a directed edge from node j to node i and $a_{j,i} = 0$, otherwise.

$$\chi_i = \frac{1}{2} \sum_{j,h \in V} (w_{i,j}w_{i,h}w_{j,h})^{1/3}. \tag{20}$$

The clustering coefficient of a node is the fraction of triangles around the node. It is defined as the fraction of the neighbors of the node that are also the neighbors of each other.

3.3.2.2. Transitivity

Transitivity of a node is similar to its clustering coefficient. However, transitivity is normalized over all nodes, while cluster coefficient for each node is normalized independently, which makes clustering coefficient biased toward nodes with low degree. Transitivity can be expressed as the ratio of triangles to triplets in the network. It is calculated by Equation (21), as suggested in Fagiolo (2007):

$$T_i = \frac{\chi_i}{\sum_{j \in V}[(d_j^{out} + d_j^{in})(d_j^{out} + d_j^{in} - 1) - 2\sum_{h \in V} a_{j,h}a_{h,j}]} \tag{21}$$

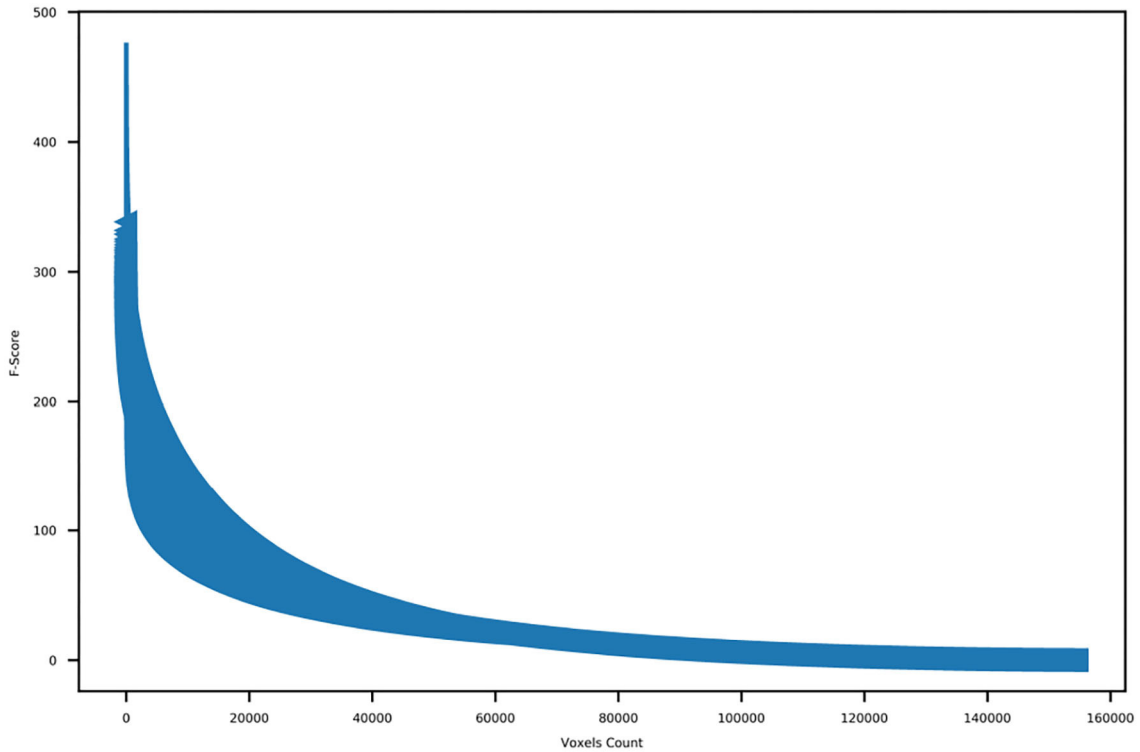


FIGURE 4 | Ordered f-scores of voxels for all subjects.

where d_j^{in} is the in-degree of node j and d_j^{out} is the out-degree of node j. χ_i is the weighted geometric mean of triangles around node i that is calculated by Equation (20). Note that $a_{h,j}a_{j,h} = 1$, if there exits an edge in both directions.

3.3.2.3. Global and Local Efficiency

The global efficiency of a brain network is a measure of its functional integration. It measures the degree of communication among the anatomical regions. Thus, it is closely related to the small-world property of a network. Formally speaking, global efficiency is defined as the average of the inverse shortest path lengths between all pairs of nodes in the brain network. Equation (22) shows how to calculate the global efficiency of a brain network, where $\varrho_{i,j}^w$ is the weighted shortest path length between two distinct nodes i and j (Rubinov and Sporns, 2010).

$$E_{global} = \frac{1}{M}\sum_{i\in V}\frac{\sum_{j\in V}(\varrho_{i,j}^w)^{-1}}{M-1} \tag{22}$$

On the other hand, the local efficiency of a network is defined as the global efficiency calculated over the neighborhood of a single node. The local efficiency is, thus, a measure of segregation rather than functional integration as it is closely related to clustering coefficient. While global efficiency is calculated for the entire network, local efficiency is calculated for each node in the network (Rubinov and Sporns, 2010).

4. EXPERIMENTS AND RESULTS

In this section, we explore the validity of the suggested dynamic brain network model and study the network properties of complex problem solving task on TOL dataset. First, we analyze the effect of the preprocessing step on the brain decoding performance of planning and execution phases of complex problem solving. Then, we investigate the validity of the dynamic functional brain network model proposed in this study. Finally, we analyze the network properties of the constructed functional brain networks for planning and execution subtasks.

4.1. Voxel Selection

First, we discarded all of the voxels located in the cerebellum anatomical regions. Then, we calculated the f-score for each one of the remaining voxels and order the obtained f-scores of the voxels. Following that, we plotted the ordered f-scores of the voxels in order to determine the appropriate number of voxels to retain. **Figure 4** shows the ordered f-scores of the voxels averaged across all subjects. It can be observed from this figure that a relatively small number of voxels is crucial for discriminating the subtasks of problem solving while the remaining voxels do not have significant information concerning the subtasks of problem solving. Based on the f-score distribution shown in **Figure 4**, we kept the 10,000 voxels with the highest

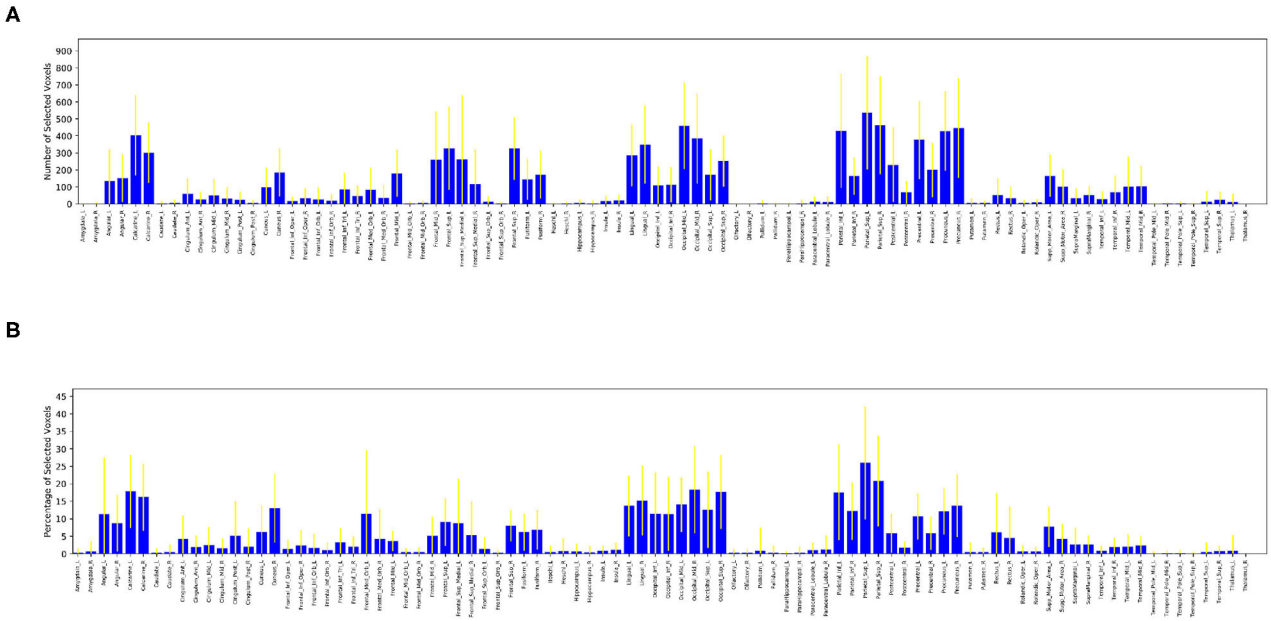


FIGURE 5 | Distribution of selected voxels across anatomical regions, measured by number of selected voxels (top) and percentage of selected voxels (bottom) from each anatomical region. A large value in both figures is an indication of relatively high activity in a particular anatomic region. **(A)** Average number of voxels selected from each anatomical region across all subjects. **(B)** Average percentage of voxels selected from each anatomical region across all subjects.

f-scores observing the elbow point, whereas we discarded the remaining ones.

After selecting 10,000 voxels with the highest f-scores of each run, we calculated the number of selected voxels contained in each one of the 90 anatomical regions. We also calculated the percentage of selected voxels to the total number of voxels located in each anatomical region. These values of selected voxels can be considered as measures of participation of an anatomic region into the complex problem solving task. The **Figure 5A** shows the average number of voxels contributed by each region across all subjects with its corresponding standard deviation, **Figure 5B** shows the average percentage of voxels contributed by each region across all subjects with its corresponding standard deviation.

It is clear from these figures that a large number of regions contribute little to no voxels, such as the amygdala, caudate, heschl gyrus, hippocampus, pallidum, putamen, temporal pole, superior temporal cortex, thalamus and parahippocampus. A small number of regions contribute a significantly large number of voxels (over 300 voxels each) during complex problem solving, such as occipital, precentral, precuneus and parietal regions.

Furthermore, **Figure 5B** ensures that there is no bias against tiny anatomical regions with small number of voxels by normalizing the number of voxels selected from each region by its total number of voxels. **Figure 5B** clearly shows that in the *left prefrontal and inferior occipital* regions a significant percentage of voxels are active during complex problem solving. Both figures also show high standard deviations across subjects, which indicates high inter-subject variability.

4.2. Interpolation

After selecting the most discriminative voxels and averaging their BOLD responses with respect to their corresponding brain anatomical regions, we employed temporal interpolation to each representative time series to increase the temporal resolution of the TOL dataset. As a result, the total number of obtained brain volumes is equal to $n + z * (n - 1)$ where n is the number of measured brain volumes of a given puzzle and z is the number of estimated brain volumes plugged between each pair of measured brain volumes. The optimal value of z is equal to 8, which is determined empirically using cross-validation to maximize the brain decoding performance. **Figure 6** shows the interpolated BOLD response of a randomly selected anatomical region from the given subjects, where the blue dots represent the measured BOLD response of the region and the orange dashes are the interpolated values. It is clear from **Figure 6** that the interpolated points using cubic spline function do not introduce sharp edges, nor do they smooth out the spikes between measured brain volumes.

Furthermore, **Figure 7** shows the single-sided amplitude spectrum of a randomly selected anatomical region from a given subject before interpolation, after interpolation and finally, after adding Gaussian noise. The figure clearly demonstrates that both interpolation and injecting Gaussian noise preserve the envelope of the signal in the frequency domain.

4.3. Gaussian Noise

In order to control the signal-to-noise ratio (SNR), we used cross-validation to choose the optimal pair of values for α_{noise} and

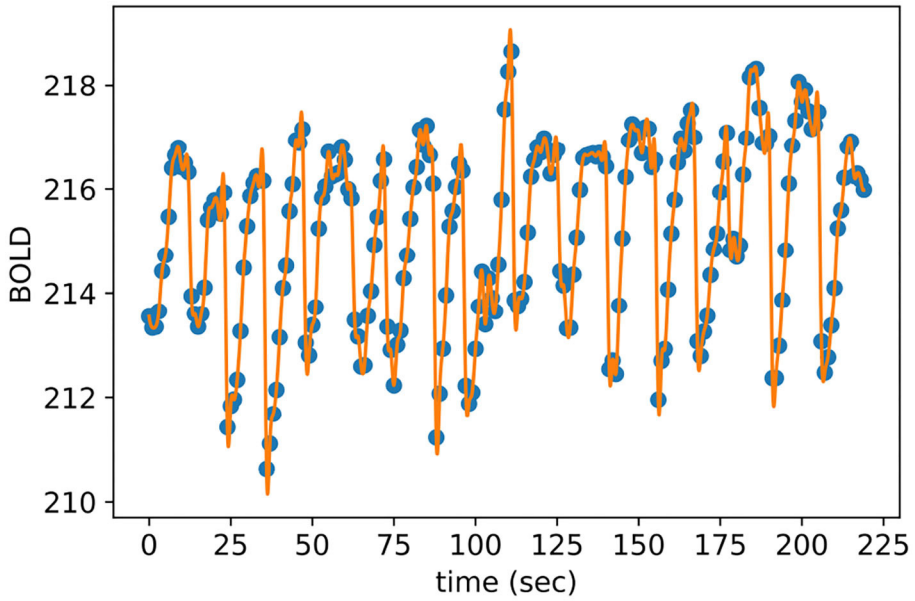


FIGURE 6 | Interpolated BOLD response of a randomly selected anatomic region.

β_{noise}, the ratios of mean and standard deviation of the added noise, respectively. As a result, the optimal values obtained are $\alpha_{noise} = 0.025$ and $\beta_{noise} = 0.075$ from the following set of values $\alpha_{noise}, \beta_{noise} \in [0.025, 0.05, 0.075, 0.1]$.

4.4. Brain Decoding With Preprocessed fMRI Data

We use brain decoding in order to investigate the validity of our proposed preprocessing steps on the TOL dataset. We aim to distinguish the two phases of complex problem solving, namely: planning and execution. At first, we used ANOVA to select the 10,000 voxels with the highest f-scores. Then we averaged the selected voxels into their corresponding anatomical regions defined by Tzourio-Mazoyer et al. (2002). Following that, we employed temporal interpolation to increase the temporal resolution of each puzzle by estimating $z = 8$ brain volumes between each pair of measured brain volumes. Finally, we added Gaussian noise in order to regularize the BOLD responses of each region to improve the generalization performance of the classifiers. We used k-fold Cross validation for each subject in all of the experiments introduced in this section, with $k = 8$. After we obtained the results, we averaged them across the different folds, then we calculated the average and standard deviation across all subjects. We used both supervised and unsupervised brain decoding methods. A linear support-vector machine (SVM) (Fan et al., 2008) was used for supervised brain decoding while k-means clustering was used for unsupervised brain decoding. The input to the decoders is formed by concatenating the values of representative time series computed per each time instant, across the anatomic regions. Considering the fact that there is a total of 90 anatomic regions, the dimension of the input vectors is 90. If there are no selected voxels in an anatomic region after the voxel selection process, the corresponding entry of the input vector becomes 0.

Table 2 shows the effect of our preprocessing pipeline on the brain decoding of complex problem solving subtasks. The first row shows the performances of brain decoding on the raw dataset without any preprocessing, simply averaging all of the voxels into their corresponding anatomical regions. While the second row shows the results of applying voxel selection then averaging the selected voxels into their anatomical regions. The third row shows the results of brain decoding after applying temporal interpolation, while the forth row shows the results after injecting the data with Gaussian noise.

From the results of the preprocessing experiments, it is observed that voxel selection improves the brain decoding performance for both supervised and unsupervised methods from 60 to 74% and from 63 to 85%, respectively. This can be attributed to the fact that voxel selection retains only the most discriminative voxels and trashes the remaining non informative ones. In addition, voxel selection manages to sparsify the representation of the data since some brain regions contribute no voxels at all; thus have no contribution to brain decoding.

The table also shows that temporal interpolation further improves the supervised brain decoding performance from 74 to 81%; this significant increase is due to increasing the number of brain volumes, thus, increasing the number of training samples for the SVM classifier. However, temporal interpolation slightly reduces the performance of unsupervised methods from 85 to 84%. This result can be partially attributed to the fact that the additional brain volumes smooth the mixture distribution, thus reducing the distinction between the two phases of problem solving, planning and execution. This is due to the method used to label the estimated brain volumes, where each estimated brain

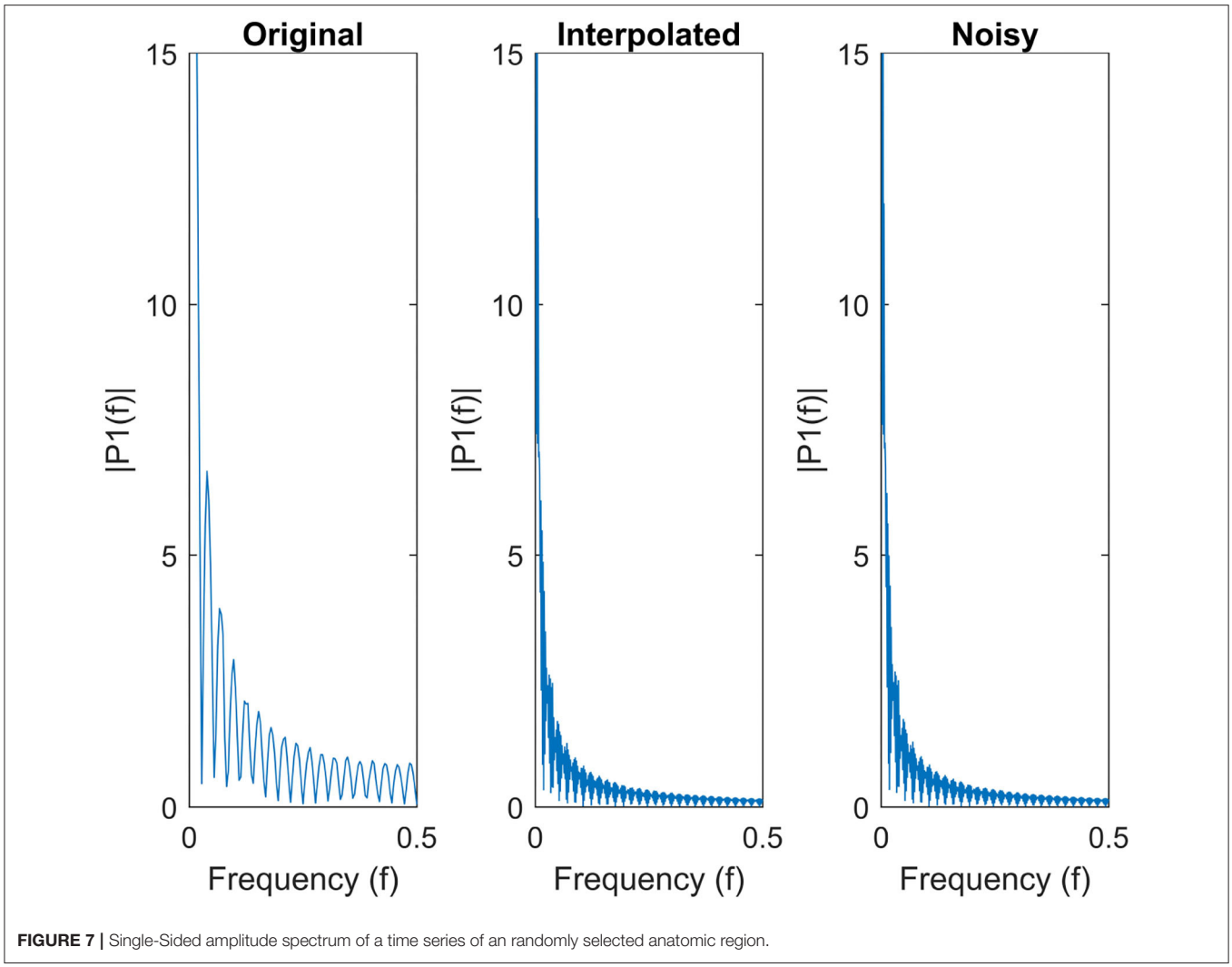


FIGURE 7 | Single-Sided amplitude spectrum of a time series of an randomly selected anatomic region.

volume is given the labels of its closest neighboring measured brain volume.

Finally, the addition of Gaussian noise slightly boosts the performance of both supervised and unsupervised methods from 81 to 82% and from 84 to 85%, respectively. The table also shows high standard deviation across subjects, which is consistent with voxel selection plots, revealing high inter-subject variability.

TABLE 2 | Decodinge performances of preprocessing pipeline after each step.

Preprocessing	SVM	*k*-Means
Raw data	0.60 ± 0.11	0.63 ± 0.09
Voxel selection	0.74 ± 0.12	**0.85** ± 0.06
Interpolation	0.81 ± 0.08	0.84 ± 0.06
Noise addition	**0.82** ± 0.08	**0.85** ± 0.06

Bold is used to indicate the best values obtained across a given column (i.e. the highest accuracy obtained).

4.5. Brain Decoding With Dynamic Brain Networks

In this section, we compare our model for building dynamic functional brain networks with some of the popular network methods proposed in the literature in terms of their brain decoding performance. Brain decoding performance can be considered as a measure of validity of the proposed brain networks. High decoding performance indicates that the constructed brain network has a good representation power of the underlying cognitive subtasks, namely planning and execution.

For this purpose, we built the dynamic brain networks, as explained in the previous sections, after having successfully applied the preprocessing pipeline. It is important to remark that each time instance has either a planning label or an execution label. While constructing the brain networks, we define a feature vector for each time instance by using the interpolated time instances (4 extra instances at each side of a measured instance). Thus, for each measured time sample, we form 4+4+1= 9 brain volumes to estimate the brain network weights. These weights represent a network among 90 anatomic regions for each measured time instance.

TABLE 3 | Braine decoding performances of proposed dynamic brain network model compared to the state of the art models, namely, pearson correlation and ridge regression.

Input to algorithm	SVM	*k*-Means
Preprocessed fMRI Data	**0.82** ± 0.08	0.85 ± 0.06
Pearson correlation	0.58 ± 0.05	0.57 ± 0.04
Ridge regression	0.56 ± 0.05	0.55 ± 0.02
Dynamic brain networks	**0.82** ± 0.10	**0.87** ± 0.06

Bold is used to indicate the best values obtained across a given column (i.e. the highest accuracy obtained).

The optimal values for learning rate $\alpha_{learning}$ and the number of epochs were chosen empirically using cross-validation, obtaining the following values, respectively $1 * 10^{-8}$ and 10. As for p, the number of neighbors used to represent each anatomical region; we chose p equal to the total number of regions, which is 90. In this way, a fully-connected brain network is obtained at each time window. However, the total number of nodes is less than 90, given that some regions have flat BOLD responses; therefore, they were pruned along with all their edges from the brain network.

We also constructed brain networks using Pearson correlation and ridge regression as proposed in Richiardi et al. (2011), Onal et al. (2015), Ertugrul et al. (2016), and Onal et al. (2017), respectively, in order to compare the performance of our methods with other works in the literature. In the case of Pearson correlation, the functional brain networks were constructed using Pearson correlation scores between each pair of brain regions (Richiardi et al., 2011; Ertugrul et al., 2016). As for the case of ridge regression, the mesh arc-weight descriptors were estimated using ridge regression in order to represent each region as a linear combination of its neighbors (Onal et al., 2015, 2017).

Since our goal is to represent the fMRI data by an informative dynamic network structure, we used generic classification/clustering methods with relatively small learning capacity in order to highlight the representative power of the constructed brain networks. For this reason, we used simple classifiers/clustering methods, such as SVM and K-means. It would be possible to improve the brain decoding performances by using methods with higher learning capacity, such as Multi-Layer Perceptrons. In this case, the dynamic network representation of the Artificial Neural Networks is expected to obtain much better performances compared to the ones reported in the paper. However, the reported performances are sufficient to show that the decoding performance of the dynamic network, which is 90 x 90 = 8,100 edges of the brain network, is compatible with the fMRI data, although it is much more compact and more informative than the raw data (185,000 voxels per brain volume).

The main advantage of representing the fMRI data by dynamic brain networks is that they are neuroscientifically interpretable and much more comprehensive compared to the voxel based representations. The constructed dynamic brain networks allow us to investigate a large variety of network properties in order to identify regions of interest, such as hubs and subgroups of densely connected brain regions with the aim of deriving neuro-scientifically valid insights into the planning and execution phases of the complex problem solving task.

Table 3 shows the brain decoding results of the aforementioned brain network construction methods compared against the results of multi-voxel pattern analysis (MVPA), which feeds the preprocessed BOLD response representing a time instant of all brain regions into a classifier. The first row shows the brain decoding results of preprocessed fMRI data, where there is no network representation at all. This data is obtained by applying voxel selection, interpolation then noise addition to the raw fMRI data. The first row of **Table 3** is the same as the last row of **Table 2**.

The second and third rows show the brain decoding performances of the networks extracted using Pearson correlation and Ridge regression methods, respectively. The Pearson Correlation data is generated using the preprocessed fMRI data, where Pearson correlation is used to generate the brain networks. The weights of the edges in the constructed brain networks are the Pearson correlation scores between the preprocessed BOLD responses of the corresponding anatomic regions (as shown in Equation 6). The Ridge Regression data is generated using the preprocessed fMRI data, where Ridge Regression is used to estimate the weights of the edges in the constructed brain networks. The weights of the edges in the constructed brain networks are estimated using Ridge regression by minimizing the cost function of Equation (9).

The last row shows the brain decoding performances of our proposed Dynamic Brain Network model.

The edge weights of the Dynamic Brain Network is computed by training an Artificial Neural Network with the preprocessed fMRI data. The nodes in the dynamic brain networks represent the brain anatomic region. The edge links of the brain networks are determined by using Equation (7). The edge weights are estimated using Artificial Neural Network as shown in Equations (10) and (11).

The inputs to SVM and K-Means in the case of Pearson Correlation, Ridge Regression and Artificial Neural Networks are the estimated weights of the brain networks. A feature vector of edge weights, with 90 x 90 = 8,100 dimension, is defined at each recorded time instance of fMRI data, as a single training/testing sample. Each feature vector has its corresponding class label, as Planning or Execution.

Table 3 clearly shows that both Pearson correlation and Ridge regression fail to construct valid brain networks that are good representatives of the underlying cognitive tasks. However, our model managed to get brain decoding results similar or slightly better than those obtained from MVPA both in the cases of supervised and unsupervised methods. This can be attributed to the challenging nature of the TOL dataset; Pearson correlation does not manage to capture the interdependencies between the anatomical regions over short time windows. While ridge regression fails to correctly estimate the mesh arc-weights as it estimates the arc-weights for each region independently of the other ones. Our proposed model, with a relatively small number of epochs, manages to obtain mesh arc-weight values that capture the activation patterns of anatomical regions and their relationships.

It is important to note that, it is possible to obtain higher brain decoding accuracy using voxel-level MVPA rather than anatomic-region-level MVPA, and by normalizing the BOLD response of individual voxels to having 0 mean and 1 standard deviation. However, we do not employ either of them in our analysis for the following reasons. Firstly, any analysis at the voxel-level comes with a very high computational cost, especially when attempting to build functional brain networks. Also, the analysis at voxel level does not allow us to perform the study roles and contributions of brain anatomic region level to the complex problem solving task, which is the main goal of this paper. Secondly, normalizing the BOLD responses of individual voxels prevents us from constructing informative, functional brain networks as this normalization distorts the information of relative activation patterns between the voxels and the brain anatomic region, which is essential in the process of building functional brain networks.

5. BRAIN NETWORK PROPERTIES

In this section, we aim to analyze the network properties of the constructed functional brain networks. We investigate the network properties for each anatomical brain region during both planning and execution subtasks in order to understand which regions are most active and which regions work together during each one of the two subtasks of complex problem solving.

Given that the constructed brain functional networks are both weighted, directed, fully-connected and contain both negative and positive weights, we preprocessed the networks before measuring their network properties. Firstly, we got rid of all the negative weights by shifting all the mesh arc-weights values by a positive quantity equal to the absolute value of the largest negative arc-weight. We then normalized the mesh arc-weights to ensure that all of the weights are within the range of $[0, 1]$. Finally, we measured the network properties on the pruned brain graph, where the brain regions (nodes) contributing no voxels (have a flat BOLD response) and all of their corresponding arc-weights (edges) were deleted from the brain graph. Thus, the networks contained less than 90 regions with their corresponding edges. We used the brain connectivity toolbox to calculate the investigated network properties (Rubinov and Sporns, 2010).

In order to measure for centrality, the number of neighbors for each anatomical region (P) was chosen to be equal to 89, which is equal to the total number of neighbors for any given node as the total number of brain anatomical regions defined by the AAL atlas (Tzourio-Mazoyer et al., 2002) after deleting the regions residing in the cerebellum equals 90. In addition, since we pruned the nodes that correspond to regions from which no voxels were selected, our constructed brain networks were weighted directed fully-connected networks. Therefore, the in-degree, out-degree and total degree of all nodes in the graph were equal to the total number of anatomical regions retained after voxel selection.

Therefore, we used node strength and node betweenness centrality to identify nodes with high centrality, which are potential hubs in the brain networks controlling the flow of information in the network. In our proposed model, the node in-strength of node i is the sum of the mesh arc-weight values, which is estimated using our proposed neural network method in order to minimize the reconstruction error of the BOLD response of anatomical region i using its neighbors. Thus, node in-strength is not used as part of our network properties analyses; we rather used node out-strength to measure the centrality of all anatomicalregions.

As for measures of segregation, quantifying the existence of subgroups within brain networks is based on densely interconnected nodes. These subgroups are commonly referred to as clusters or modules. The existence of such clusters in functional brain networks is a sign of interdependence among the nodes forming the cluster. Therefore, clustering coefficient, transitivity and local efficiency were measured in order to identify potential clusters with dense interconnections in the brain networks.

5.1. Planning and Execution Brain Networks

In this section, we discuss the network properties of the planning and execution networks. For each aforementioned network metric, we ranked the brain regions in descending order according to their score on that network measure for all subjects across all runs. Then, we retained the 10 anatomical regions with the highest scores. Following that, we measured the frequency of occurrence of each brain region among the top 10 regions across all runs in order to identify the shared regions and patterns across all subjects for both planning and execution subtasks. The results of the analysis are shown in tables: **Table 4** shows the brain regions that have high scores for the reported network properties during planning subtask, and **Table 5** shows the brain regions that have high scores during execution subtask. There are a number of processes taking place during planning and execution. Plan generation involves a series of recursive events including 1) problem encoding; 2) decision-making in order to decide which ball to move and where to move it; 3) mental imagery to imagine the ball moving, and 4) working memory to maintain the intermediate steps as well as the move number. During plan execution, there is 1) retrieval of the steps from memory; 2) confirming the correct steps are being performed, and 3) the motor execution of those steps. As the results demonstrate, the networks for planning and execution are overlapping. These results are similar to the activation results reported in Newman et al. (2009) in that the regions that were found to be active during the task are also regions that are most prominently found with the highest network measures. These regions include the right and left middle frontal gyrus, anterior cingulate cortex, precentral cortex, and superior parietal cortex.

Previous work has suggested that the regions found in the current study to show high network measures are directly related to the sub-tasks associated with TOL performance. For example, both the left and right prefrontal cortex have been found to be involved in the TOL task, with the two regions performing distinguishable functions. The right prefrontal cortex is involved in constructing the plan for solving the TOL problem while the left prefrontal cortex is involved in supervising the execution of that plan (Newman et al., 2003, 2009). The anterior cingulate has been linked to error detection and is particularly

TABLE 4 | Planning: Anatomical regions with the highest network measures across subjects, regions are painted if they overlap with execution.

Transitivity	Local efficiency	Clustering Coefficient	Betweenness	Out-strength
Angular	Calcarine	Calcarine	Cueneus R	Cueneus R
Calcarine	Cuneus	Cuneus	Frontal Sup L	Frontal Sup L
Cingulum Ant	Frontal Mid R	Frontal Mid R	Fusiform R	Fusiform R
Cingulum Mid	Frontal Sup	Frontal Sup	Paracentral Lobule L	Paracentral Lobule L
Cuneus	Fusiform	Fusiform	Parietal Sup R	Supp Motor Area R
Frontal Inf Oper L	Occipital Inf R	Occipital Inf R	Precuneus L	Temporal Inf R
	Precentral	Parietal Sup R	Supp Motor Area R	Temporal Mid R
	Supp Motor Area R	Precentral	Temporal Inf R	
	Temporal Inf R	Supp Motor Area R	Temporal Mid R	
		Temporal Inf R		

TABLE 5 | Execution: Anatomical regions with the highest network measures across subjects, regions are painted if they overlap with planning.

Transitivity	Local efficiency	Clustering coefficient	Betweenness	Out-strength
Angular	Calcarine	Calcarine	Cueneus R	Cueneus R
Calcarine	Cuneus	Cuneus	Frontal Sup L	Frontal Sup L
Cingulum Ant	Frontal Mid R	Frontal Mid R	Fusiform R	Fusiform R
Cingulum Mid	Frontal Sup	Frontal Sup	Paracentral Lobule L	Paracentral Lobule L
Cuneus	Fusiform	Fusiform	Parietal Sup R	Supp Motor Area R
Frontal Inf Oper L	Occipital Inf R	Occipital Inf R	Precuneus L	Temporal Inf R
	Precentral	Parietal Sup R	Supp Motor Area R	Temporal Mid R
	Supp Motor Area R	Precentral	Temporal Inf R	
	Temporal Inf R	Supp Motor Area R	Temporal Mid R	
		Temporal Inf R		

involved in the TOL when the number of moves is higher, or the problem difficulty is manipulated. The right superior parietal cortex and precentral cortex have been linked to visuo-spatial attention necessary for planning (Newman et al., 2003), and the left parietal cortex has been linked to visuo-spatial working memory processing (Newman et al., 2003). The overlap between the regions with the highest network measures and those that have been linked to the task is an important feature and is not due to the voxel selection process. Many regions that passed threshold were not in the top ranked list of network measures. For example, the basal ganglia, including the caudate has been found in previous studies to be involved in TOL performance (Dagher et al., 1999; Rowe et al., 2001; Beauchamp et al., 2003; Van den Heuvel et al., 2003; Newman et al., 2009); however, the region appears to not be an important network hub. **Figures 8A,B** visualize the reported brain regions in **Tables 4**, **5**, respectively, using Brain Net Viewer (Xia et al., 2013). In **Figures 8A,B**, the color of the node (brain region) implies the following: red indicates that the region has high transitivity, clustering coefficient or local efficiency. Green indicates that the node has high node centrality measured by node out-strength and node betweenness. As for blue, it shows the nodes that have high node centrality and is part of a subgroup of densely interconnected regions.

5.2. Differences Between Planning and Execution Networks

In this section, we explore the network differences between planning and execution by calculating the difference between the network property scores for planning and execution for each run. To achieve that, we took the difference between the network property scores for brain anatomical regions during planning and the network property scores for brain anatomical regions during execution for each run. Then, we counted the frequency of times a given anatomical region is more active during planning than execution and vice-versa in order to identify consistent patterns of the disagreements between planning brain networks and execution brain networks across all subjects. Results showed, generally, that the network measures were higher for planning than execution. This, too, mirrors the findings from Newman et al. (2009) in which planning resulted in greater activation than execution.

Node out-strength is a measure of how connected the node is to other nodes in the network. **Figures 9A,B** visualize the brain regions with higher node out-strength during planning and during execution, respectively. Planning showed greater out-strength than execution in the following regions: occipital regions (calcarine, cuneus), parietal regions (bilateral superior parietal cortex and precunues), the right superior frontal cortex,

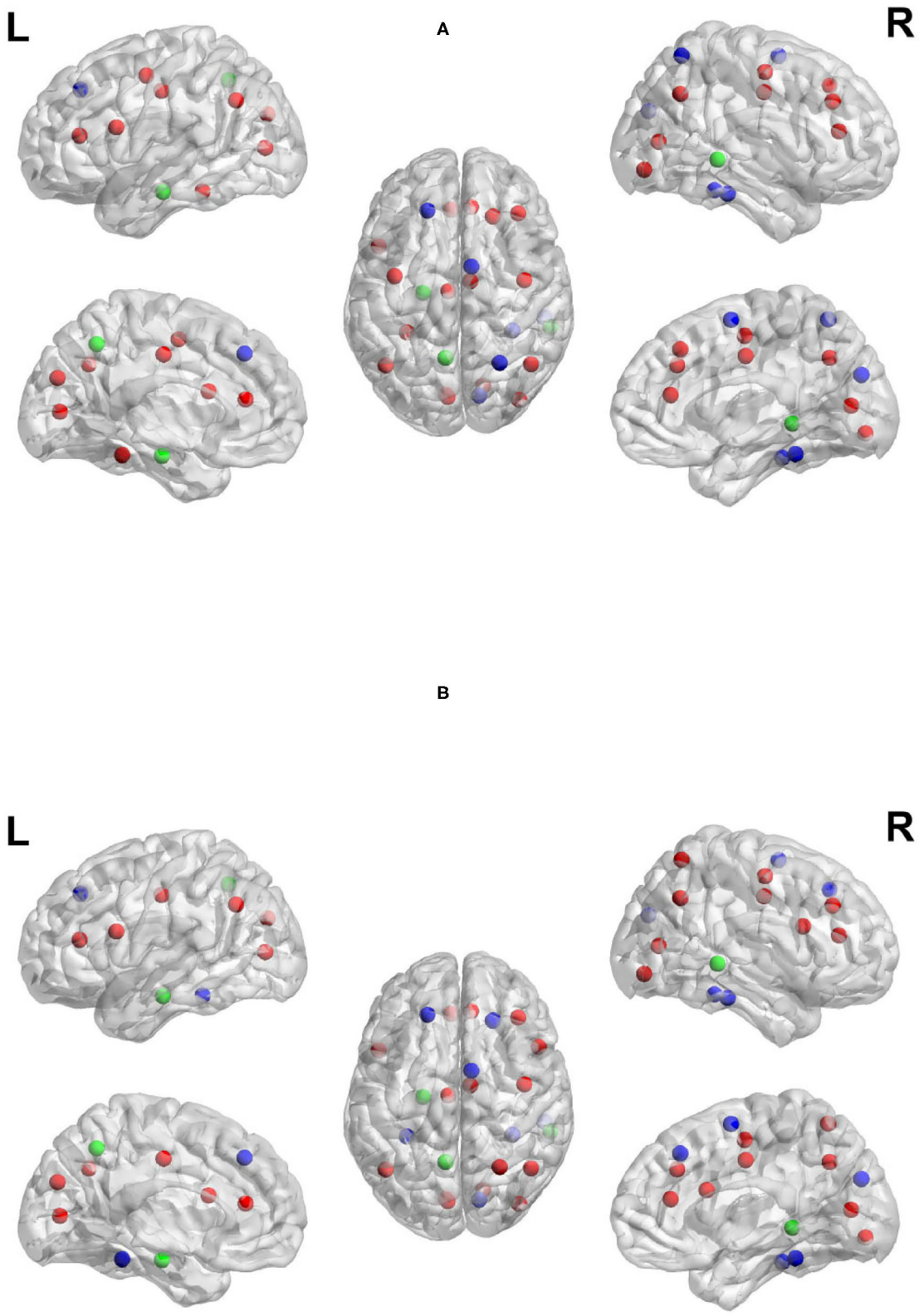


FIGURE 8 | Regions with the highest network measures for planning brain network **(A)** and execution brain network **(B)**.

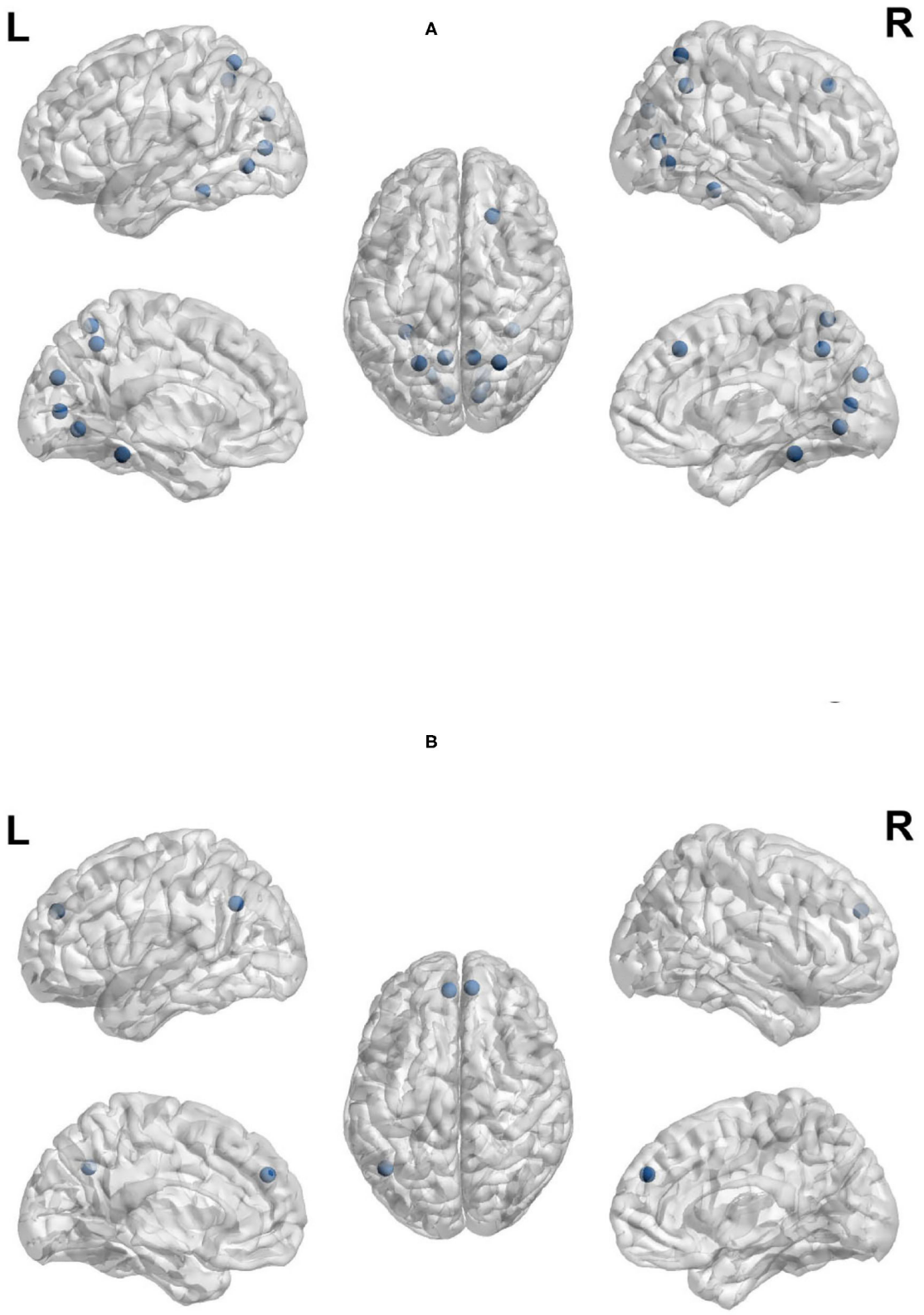


FIGURE 9 | Anatomical regions with higher node out-strength during planning (Top) and during execution (Bottom). **(A)** Anatomical regions with higher node out-strength during planning. **(B)** Anatomical regions with higher node out-strength during execution.

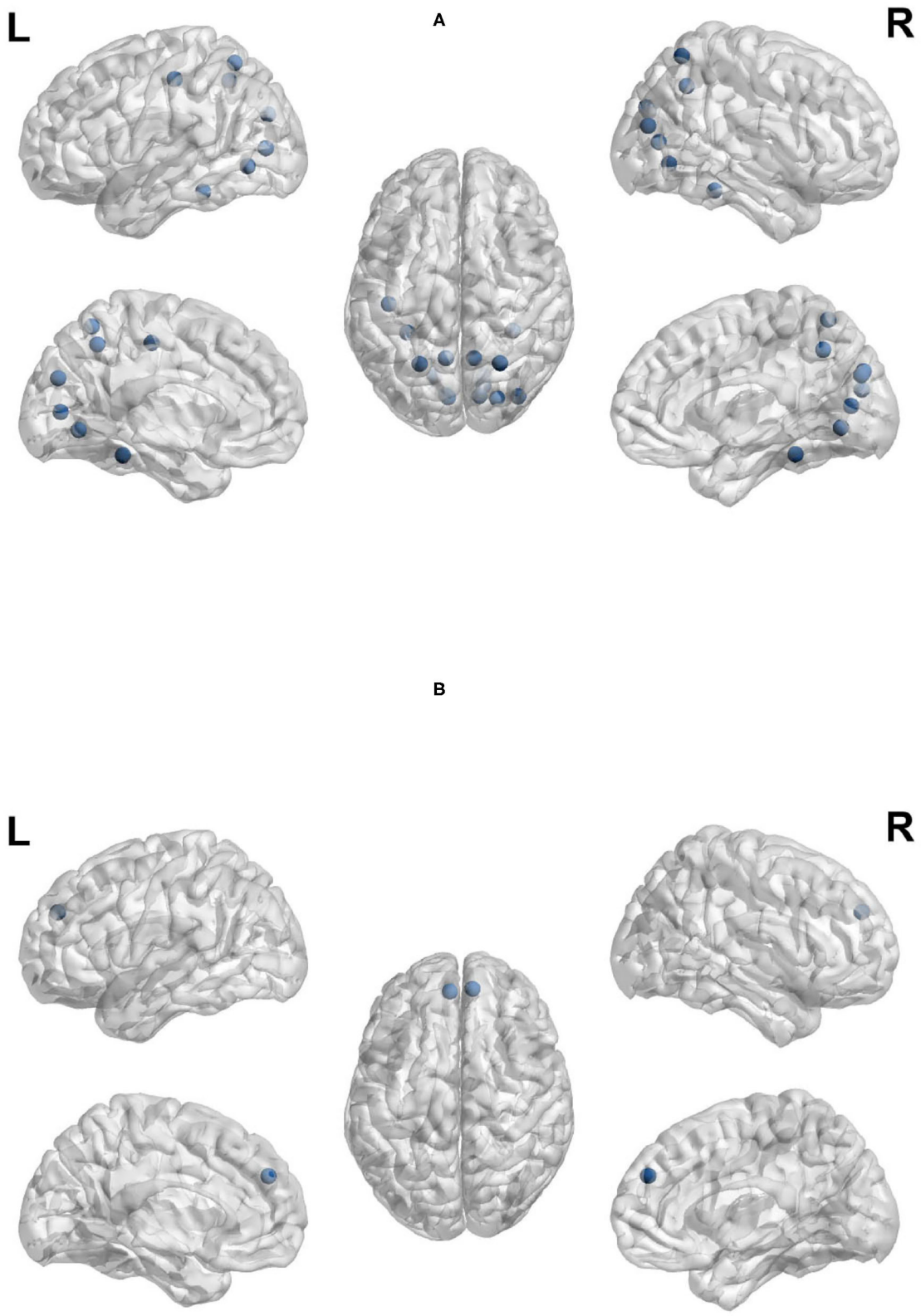


FIGURE 10 | Anatomical regions with higher node betweenness during planning (Top) and during execution (Bottom). **(A)** Anatomical regions with higher node betweenness during planning. **(B)** Anatomical regions with higher node betweenness during exection.

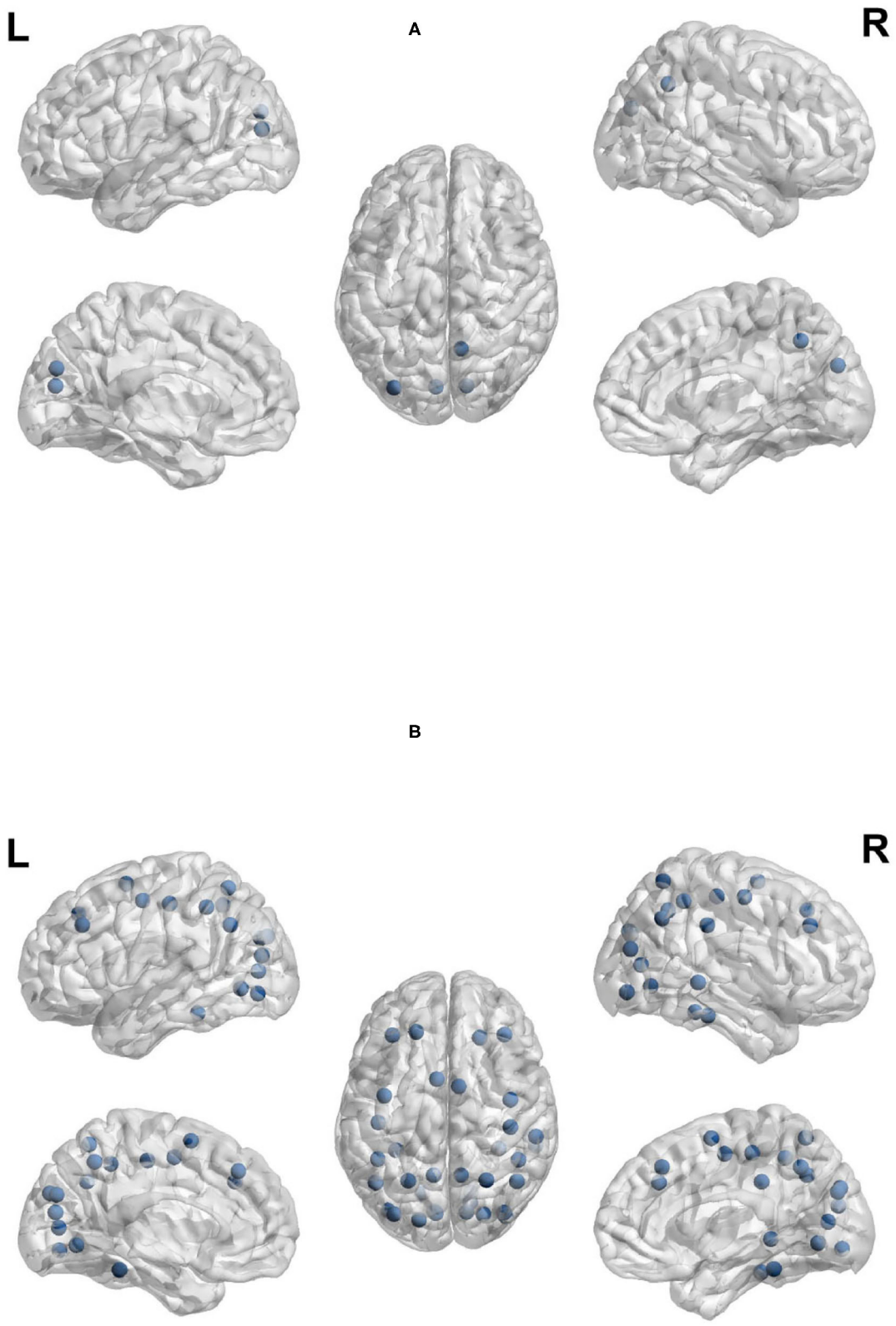


FIGURE 11 | Anatomical regions with higher local efficiency and clustering coefficient (Top) and higher transitivity (Bottom) during planning. **(A)** Anatomical regions with higher local efficiency and clustering coefficient during planning. **(B)** Anatomical regions with higher transitivity during planning.

and inferior occipito-temporal regions (fusiform and lingual gyri). The left angular gyrus and bilateral medial superior frontal cortex showed greater out-strength for execution. As for node betweenness, **Figures 10A,B** visualize the brain regions with higher betweenness during planning and during execution, respectively. The following brain regions had higher node betweenness during planning than execution: occipital regions (calcarine, cuneus, right middle, right superior); inferior occipito-temporal (fusiform, lingual); parietal (bilateral superior parietal, left postcentral, precuneus). Bilateral medial superior frontal had higher node betweenness during execution than planning.

These results suggest that there is greater information flow during planning than execution. This matches our expectations. Planning is more computationally demanding than execution. Again, during planning, participants must explore the problem space, which requires generating and manipulating a mental representation of the problem. The regions that show greater information flow during planning are all regions involved in that generation and manipulation, particularly parietal, occipital and inferior occipito-temporal. On the other hand, execution requires recall of the plan generated and stored and therefore, greater information flow from frontal regions related to memory retrieval is observed.

Clustering coefficient, local efficiency and transitivity are measures of segregation that aim to identify sub-networks. **Figure 11A** visualizes the brain regions with higher local efficiency and higher clustering coefficient during planning phase compared to execution phase. While **Figure 11B** visualizes the brain regions with higher transitivity during planning than during execution phase. Each of these measures was larger for planning than execution, with no regions showing larger measures for execution.

The regions that showed a higher clustering coefficient in planning included: the cuneus, left middle occipital cortex, and right precuneus. Local efficiency was higher in a similar set of regions (the cuneus, left middle occipital cortex, and right precuneus). The clustering coefficient and local efficiency identified a visual-spatial sub-network that is more strongly connected during planning. Transitivity identified an overlapping but more extensive set of regions that included: bilateral angular gyrus, calcarine sulcus, cuneus, bilateral middle frontal cortex, bilaterial superior frontal cortex, bilateral fusiform and lingual gyri, bilateral occipital cortex, bilatral superior parietal cortex, postcentral and precentral cortex, precuneus, supplementary motor area, right supramarginal gyrus, and right inferior and middle temporal cortex.

5.3. Global Efficiency

Since global efficiency is measured over the entire brain network, not for a given node in the network, we measured the global efficiency for all planning and execution networks within all runs across subjects. Then, global efficiency of planning is compared against that of execution. Results show that the majority of runs had higher global efficiency scores during planning than execution; 43 out of 72 runs had higher global efficiency during planning than execution. Furthermore, **Table 6** shows the number of runs where global efficiency was higher during planning and during execution across all subjects for all 4 runs of each subject. The first column shows the number of subjects that had a higher global efficiency score during planning than during execution. The second column shows the number of subjects that had a higher global efficiency score during execution than during planning.

TABLE 6 | Global efficiency.

Run number	Planning	Execution
1	15	3
2	9	9
3	10	8
4	9	9

Although there was no significant difference in global efficiency between planning and execution, from the table, it is clear that the majority of subjects had a higher global efficiency for planning for the first runs. Some subjects switched from having higher global efficiency during planning to having higher global efficiency during execution. A potential explanation for this change across runs is switching from pre-planning to online planning or planning intermixed with execution. Although there is a dedicated planning phase in the current study, that does not mean that planning is not taking place during execution. In fact, it has been debated as to whether efficient pre-planning is possible in the TOL or whether TOL performance is controlled by online planning (Kafer and Hunter, 1997; Phillips, 1999, 2001; Unterrainer et al., 2004). According to Phillips (1999, 2001), pre-planning the entire sequence is not natural, but that people instead plan the beginning sequence of moves and then intersperse planning and execution. If this is the case, then it may be expected that some participants will switch to online planning. This intermixing of planning and execution is also likely to impact the performance of the machine learning algorithms to detect planning and execution phases.

The relationship between global efficiency and behavioral performance was examined. Global efficiency was found to be positively correlated with the mean number of extra moves (a measure of error) during problem-solving (for execution $r = 0.73$, $p = 0.0006$). Previous studies have shown a relationship between global efficiency and task performance (Stanley et al., 2015).

This suggests that the variance in global efficiency is indicative of individual differences in neural processing and further suggests that the changes in global efficiency across runs are also likely indicative of changes in neural processing related to changing strategy. Further research using a larger sample is necessary to explore this hypothesis.

6. CONCLUSION

In this paper, we propose a new computational method to estimate dynamic functional brain networks from the fMRI signal recorded during a complex problem solving task. Our model recognizes the two phases of complex problem solving

with more than 80% accuracy, indicating the representation power of the suggested dynamic brain network model. We study the properties of the constructed brain networks during planning and execution phases in order to identify essential anatomic regions in the brain networks related to problem solving. We investigate the potential hubs and densely connected clusters. Furthermore, we compare the network structure of the estimated dynamic brain networks for planning and execution tasks.

There are some limitations to the study. Although the primary aim of this study was to demonstrate the feasibility of the methods, the sample size is somewhat small, making the interpretation of the results difficult. Second, a goal of this method is to identify brain states that are interspersed with each other. In the current study, planning was expected to occur both prior to execution as well as during execution; therefore, planning states are interspersed within the execution phase. The temporal sampling rate of the fMRI data may be a limiting factor. Alternatively, the sluggish and blurred underlying hemodynamic response may be the factor preventing the ability to detect brain states. We plan to explore this factor in future work.

DATA AVAILABILITY STATEMENT

The dataset used in this study and the code required to reproduce our results can be found at https://osf.io/krch2/?view_only=df8aaf1f4fa46129a69f89486f65a83.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Indiana University Institutional Review Board. The patients/participants provided their written informed consent to participate in this study. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. All individual participants in the study signed an informed, written consent documents approved by the Indiana University Institutional Review Board.

AUTHOR CONTRIBUTIONS

AA proposed and implemented the computational model with support from OE and BK. Performed the experiments with support from BK and OE. Prepared the visualizations with support from BK. Wrote the manuscript with support from OE, SN, and FY. BK proposed the idea for building brain networks using neural networks. SN provided the neuroscientific interpretations of the results of the proposed computational model, designed the Tower of London experiment procedure and collected the corresponding fMRI recordings. FY supervised the project, conceived the study and was in charge of the overall direction and planning. All authors provided critical feedback and helped shape the research, analysis and manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by TUBITAK (Scientific and Technological Research Council of Turkey) under grant No: 116E091 as well as the Indiana METACyt Initiative of Indiana University, funded in part through a major grant from the Lilly Endowment, Inc. AA, OE, BK, and FY received funding from TUBITAK (Scientific and Technological Research Council of Turkey) under the grant no: 116E091. SN received research funding from the Indiana METACyt Initiative of Indiana University and a major grant from the Lilly Endowment, Inc. The authors declare that this study received funding from the Lilly Endowment Inc. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

ACKNOWLEDGMENTS

We thank Gonul Gunal Degirmendereli for her contributions to this manuscript and her help with data analysis of the TOL experiment procedure.

REFERENCES

Afrasiyabi, A., Onal, I., and Yarman Vural, F. T. (2016). "A sparse temporal mesh model for brain decoding," in *Cognitive Informatics Cognitive Computing (ICCI* CC), 2016 IEEE 15th International Conference* (Palo Alto, CA: IEEE), 198–206.

Albert, D., and Steinberg, L. (2011). Age differences in strategic planning as indexed by the tower of london. *Child Dev.* 82, 1501–1517. doi: 10.1111/j.1467-8624.2011.01613.x

Alchihabi, A., Kivilicim, B. B., Newman, S. D., and Yarman Vural, F. T. (2018). "A dynamic network representation of fmri for modeling and analyzing the problem solving task," in *Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium* (Washington, DC: IEEE), 114–117.

Bassett, D. S., and Bullmore, E. (2006). Small-world brain networks. *Neuroscientist* 12, 512–523. doi: 10.1177/1073858406293182

Beauchamp, M., Dagher, A., Aston, J., and Doyon, J. (2003). Dynamic functional changes associated with cognitive skill learning of an adapted version of the tower of london task. *Neuroimage* 20, 1649–1660. doi: 10.1016/j.neuroimage.2003.07.003

Boghi, A., Rasetti, R., Avidano, F., Manzone, C., Orsi, L., D'agata, F., et al. (2006). The effect of gender on planning: an fmri study using the tower of london task. *Neuroimage* 33, 999–1010. doi: 10.1016/j.neuroimage.2006.07.022

Brandes, U. (2001). A faster algorithm for betweenness centrality. *J. Math. Sociol.* 25, 163–177. doi: 10.1080/0022250X.2001.9990249

Chang, Y.-K., Tsai, C.-L., Hung, T.-M., So, E. C., Chen, F.-T., and Etnier, J. L. (2011). Effects of acute exercise on executive function: a study with a tower of london task. *J. Sport Exerc. Psychol.* 33, 847–865. doi: 10.1123/jsep.33.6.847

Cochran, W. T., Cooley, J. W., Favin, D. L., Helms, H. D., Kaenel, R. A., Lang, W. W., et al. (1967). What is the fast fourier transform? *Proc. IEEE* 55, 1664–1674. doi: 10.1109/PROC.1967.5957

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/S1053-8119(03)00049-1

Dagher, A., Owen, A. M., Boecker, H., and Brooks, D. J. (1999). Mapping the network for planning: a correlational pet activation study with the tower of london task. *Brain* 122, 1973–1987. doi: 10.1093/brain/122.10.1973

Desco, M., Navas-Sanchez, F. J., Sanchez-González, J., Reig, S., Robles, O., Franco, C., et al. (2011). Mathematically gifted adolescents use more extensive and more bilateral areas of the fronto-parietal network than controls during executive functioning and fluid reasoning tasks. *Neuroimage* 57, 281–292. doi: 10.1016/j.neuroimage.2011.03.063

Ekman, M., Derrfuss, J., Tittgemeyer, M., and Fiebach, C. J. (2012). Predicting errors from reconfiguration patterns in human brain networks. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16714–16719. doi: 10.1073/pnas.1207523109

Ertugrul, I. O., Ozay, M., and Yarman Vural, F. T. (2016). Hierarchical multi-resolution mesh networks for brain decoding. *Brain Imaging Behav.* 12:1067–1083. doi: 10.1007/s11682-017-9774-z

Fagiolo, G. (2007). Clustering in complex directed networks. *Phys. Rev. E* 76, 026107. doi: 10.1103/PhysRevE.76.026107

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). Liblinear: a library for large linear classification. *J. Mach. Learn. Res.* 9, 1871–1874. doi: 10.1145/1390681.1442794

Fırat, O., Özay, M., Önal, I., Öztekin, İ., and Yarman Vural, F. T. (2013). "Functional mesh learning for pattern analysis of cognitive processes," in *Cognitive Informatics Cognitive Computing (ICCI* CC)*, *2013 12th IEEE International Conference* (New York, NY: IEEE), 161–167.

Frigo, M., and Johnson, S. G. (1998). "Fftw: an adaptive software architecture for the fft," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, Vol. 3* (Seattle, WA: IEEE), 1381–1384.

Goethals, I., Audenaert, K., Jacobs, F., Van de Wiele, C., Ham, H., Pyck, H., et al. (2005). Blunted prefrontal perfusion in depressed patients performing the tower of london task. *Psychiatry Res. Neuroimaging* 139, 31–40. doi: 10.1016/j.pscychresns.2004.09.007

Kafer, K., and Hunter, M. (1997). On testing the face validity of planning/problem-solving tasks in a normal population. *J. Int. Neuropsychol. Soc.* 3, 108–119. doi: 10.1017/S1355617797001082

Kaller, C. P., Heinze, K., Mader, I., Unterrainer, J. M., Rahm, B., Weiller, C., et al. (2012). Linking planning performance and gray matter density in mid-dorsolateral prefrontal cortex: moderating effects of age and sex. *Neuroimage* 63, 1454–1463. doi: 10.1016/j.neuroimage.2012.08.032

Kivilcim, B. B., Ertugrul, I. O., and Yarman Vural, F. T. (2018). "Modeling brain networks with artificial neural networks," in *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities* (Granada: Springer), 43–53.

Kurmukov, A., Ananyeva, M., Dodonova, Y., Gutman, B., Faskowitz, J., Jahanshad, N., et al. (2017). "Classifying phenotypes based on the community structure of human brain networks," in *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics* (Québec City: Springer), 3–11.

Lee, H., Lee, D. S., Kang, H., Kim, B.-N., and Chung, M. K. (2011). Sparse brain network recovery under compressed sensing. *IEEE Trans. Med. Imaging* 30, 1154–1165. doi: 10.1109/TMI.2011.2140380

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Stat. Sci.* 23:439–464. doi: 10.1214/09-STS282

Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., et al. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* 30, 9477–9487. doi: 10.1523/JNEUROSCI.0333-10.2010

MacAllister, W. S., Bender, H. A., Whitman, L., Welsh, A., Keller, S., Granader, Y., et al. (2012). Assessment of executive functioning in childhood epilepsy: the tower of london and brief. *Child Neuropsychol.* 18, 404–415. doi: 10.1080/09297049.2011.613812

Matsuoka, K. (1992). Noise injection into inputs in back-propagation learning. *IEEE Trans. Syst. Man. Cybern.* 22, 436–440. doi: 10.1109/21.155944

McKinley, S., and Levine, M. (1998). Cubic spline interpolation. *Coll. Redwoods* 45, 1049–1060.

Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn. Sci.* 15, 483–506. doi: 10.1016/j.tics.2011.08.003

Newell, A., Shaw, J. C., and Simon, H. A. (1957). *Elements of a theory of human problem solving*. Technical report, RAND CORP SANTA MONICA CALIF.

Newman, S. D., Carpenter, P. A., Varma, S., and Just, M. A. (2003). Frontal and parietal participation in problem solving in the tower of london: fmri and computational modeling of planning and high-level perception. *Neuropsychologia* 41, 1668–1682. doi: 10.1016/S0028-3932(03)00091-5

Newman, S. D., Greco, J. A., and Lee, D. (2009). An fmri study of the tower of london: a look at problem structure differences. *Brain Res.* 1286:123–132. doi: 10.1016/j.brainres.2009.06.031

Onal, I., Ozay, M., Mizrak, E., Oztekin, I., and Vural, F. T. Y. (2017). A new representation of fmri signal by a set of local meshes for brain decoding. *IEEE Trans. Signal Inf. Proc. Over Netw.* 3, 683–694. doi: 10.1109/TSIPN.2017.2679491

Onal, I., Ozay, M., and Vural, F. T. Y. (2015). "Modeling voxel connectivity for brain decoding," in *Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop* (Stanford, CA: IEEE), 5–8.

Park, H.-J., and Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science* 342, 1238411. doi: 10.1126/science.1238411

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fmri: a tutorial overview. *Neuroimage* 45, S199–S209. doi: 10.1016/j.neuroimage.2008.11.007

Phillips, F. (2001). A research note on accounting students' epistemological beliefs, study strategies, and unstructured problem-solving performance. *Issues Account. Educ.* 16, 21–39. doi: 10.2308/iace.2001.16.1.21

Phillips, L. H. (1999). The role of memory in the tower of london task. *Memory* 7, 209–231. doi: 10.1080/741944066

Power, J. D., Fair, D. A., Schlaggar, B. L., and Petersen, S. E. (2010). The development of human functional brain networks. *Neuron* 67, 735–748. doi: 10.1016/j.neuron.2010.08.017

Rasser, P. E., Johnston, P., Lagopoulos, J., Ward, P. B., Schall, U., Thienel, R., et al. (2005). Functional mri bold response to tower of london performance of first-episode schizophrenia patients using cortical pattern matching. *Neuroimage* 26, 941–951. doi: 10.1016/j.neuroimage.2004.11.054

Reed, R., Oh, S., and Marks, R. (1992). "Regularization using jittered training data," in *Neural Networks, 1992. IJCNN, International Joint Conference, Vol. 3* (Baltimore, MD: IEEE), 147–152.

Rektorova, I., Srovnalova, H., Kubikova, R., and Prasek, J. (2008). Striatal dopamine transporter imaging correlates with depressive symptoms and tower of london task performance in parkinson's disease. *Mov. Disord.* 23, 1580–1587. doi: 10.1002/mds.22158

Richiardi, J., Achard, S., Bunke, H., and Van De Ville, D. (2013). Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience. *IEEE Signal Proc. Mag.* 30, 58–70. doi: 10.1109/MSP.2012.2233865

Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., and Van De Ville, D. (2011). Decoding brain states from fmri connectivity graphs. *Neuroimage* 56, 616–626. doi: 10.1016/j.neuroimage.2010.05.081

Rowe, J., Owen, A., Johnsrude, I., and Passingham, R. (2001). Imaging the mental components of a planning task. *Neuropsychologia* 39, 315–327. doi: 10.1016/S0028-3932(00)00109-3

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Shallice, T. (1982). Specific impairments of planning. *Phil. Trans. R. Soc. Lond. B* 298, 199–209. doi: 10.1098/rstb.1982.0082

Shirer, W., Ryali, S., Rykhlevskaia, E., Menon, V., and Greicius, M. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cereb. Cortex* 22, 158–165. doi: 10.1093/cercor/bhr099

Simon, H. A., and Newell, A. (1971). Human problem solving: the state of the theory in 1970. *Am. Psychol.* 26, 145. doi: 10.1037/h0030806

Stanley, M. L., Simpson, S. L., Dagenbach, D., Lyday, R. G., Burdette, J. H., and Laurienti, P. J. (2015). Changes in brain network efficiency and working memory performance in aging. *PLoS ONE* 10:e0123950. doi: 10.1371/journal.pone.0123950

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Unterrainer, J. M., Rahm, B., Kaller, C. P., Leonhart, R., Quiske, K., Hoppe-Seyler, K., et al. (2004). Planning abilities and the tower of london: is this task measuring a discrete cognitive function? *J. Clin. Exp. Neuropsychol.* 26, 846–856. doi: 10.1080/13803390490509574

Unterrainer, J. M., Ruff, C. C., Rahm, B., Kaller, C. P., Spreer, J., Schwarzwald, R., et al. (2005). The influence of sex differences and individual task performance on brain activation during planning. *Neuroimage* 24, 586–590. doi: 10.1016/j.neuroimage.2004.09.020

Van den Heuvel, O. A., Groenewegen, H. J., Barkhof, F., Lazeron, R. H., van Dyck, R., and Veltman, D. J. (2003). Frontostriatal system in planning complexity: a parametric functional magnetic resonance version of tower of london task. *Neuroimage* 18, 367–374. doi: 10.1016/S1053-8119(02)00010-1

Xia, M., Wang, J., and He, Y. (2013). Brainnet viewer: a network visualization tool for human brain connectomics. *PLoS ONE* 8:e68910. doi: 10.1371/journal.pone.0068910

Zook, N., Welsh, M. C., and Ewing, V. (2006). Performance of healthy, older adults on the tower of london revised: associations with verbal and nonverbal abilities. *Aging Neuropsychol. Cogn*. 13, 1–19. doi: 10.1080/13825580490904183

Zook, N. A., Davalos, D. B., DeLosh, E. L., and Davis, H. P. (2004). Working memory, inhibition, and fluid intelligence as predictors of performance on tower of hanoi and london tasks. *Brain Cogn*. 56, 286–292. doi: 10.1016/j.bandc.2004.07.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 14 December 2021
doi: 10.3389/fninf.2021.744355

Recognition of Electroencephalography-Related Features of Neuronal Network Organization in Patients With Schizophrenia Using the Generalized Choquet Integrals

*Małgorzata Plechawska-Wójcik[1], Paweł Karczmarek[1], Paweł Krukow[2], Monika Kaczorowska[1], Mikhail Tokovarov[1] and Kamil Jonak[1,2]**

[1] Department of Computer Science, Lublin University of Technology, Lublin, Poland, [2] Department of Clinical Neuropsychiatry, Medical University of Lublin, Lublin, Poland

OPEN ACCESS

Edited by:
Sharlene D. Newman,
University of Alabama, United States

Reviewed by:
Chainarong Amornbunchornvej,
National Electronics and Computer Technology Center, Thailand
Ruxandra Stoean,
University of Craiova, Romania

***Correspondence:**
Kamil Jonak
k.jonak@pollub.pl

Received: 20 July 2021
Accepted: 09 November 2021
Published: 14 December 2021

Citation:
Plechawska-Wójcik M, Karczmarek P, Krukow P, Kaczorowska M, Tokovarov M and Jonak K (2021) Recognition of Electroencephalography-Related Features of Neuronal Network Organization in Patients With Schizophrenia Using the Generalized Choquet Integrals. Front. Neuroinform. 15:744355. doi: 10.3389/fninf.2021.744355

In this study, we focused on the verification of suitable aggregation operators enabling accurate differentiation of selected neurophysiological features extracted from resting-state electroencephalographic recordings of patients who were diagnosed with schizophrenia (SZ) or healthy controls (HC). We built the Choquet integral-based operators using traditional classification results as an input to the procedure of establishing the fuzzy measure densities. The dataset applied in the study was a collection of variables characterizing the organization of the neural networks computed using the minimum spanning tree (MST) algorithms obtained from signal-spaced functional connectivity indicators and calculated separately for predefined frequency bands using classical linear Granger causality (GC) measure. In the series of numerical experiments, we reported the results of classification obtained using numerous generalizations of the Choquet integral and other aggregation functions, which were tested to find the most appropriate ones. The obtained results demonstrate that the classification accuracy can be increased by 1.81% using the extended versions of the Choquet integral called in the literature, namely, generalized Choquet integral or pre-aggregation operators.

Keywords: schizophrenia, extended Choquet integral, classifiers, aggregation, Sugeno fuzzy measure

INTRODUCTION

Mental illnesses are usually long-lasting conditions associated with great psychological suffering, the substantially limited possibility of independent functioning, and social development. Among them, schizophrenia (SZ) is one of the most severe forms of mental health disorder with the complex and multidimensional clinical picture. The onset of SZ occurs most often in adolescence

or early adulthood commonly has a slow and hidden course consisting of gradual augmenting of the so-called negative syndromes, i.e., loss of interests, affective blunting, reduced initiative, and social isolation, and more or less delayed phase of active psychotic exacerbation characterized by the presence of delusions, i.e., incorrect judgments of reality and the behavior of other people, as well as hallucinations, i.e., incorrect sensory impressions, most often in the auditory form (Rosen et al., 1984; Heilbronner et al., 2016). In addition to the negative and positive symptoms, there are also various cognitive disorders including disturbances in the course of thinking and deficits in specific cognitive domains, such as attention, memory, cognitive speed, language, and communication, and difficulties with adapting to new circumstances and problem-solving (Szöke et al., 2008; Krukow et al., 2017; Green et al., 2019). It should be noted that long-term pharmacological treatment of the disease is the main form of therapeutic intervention focused mainly on psychotic syndromes. However, even when modern methods of treatment are applied, distortions of cognitive processes improve to a much lesser extent, often causing lifelong constraints in achieving full independence (Keefe, 2019). Personal, social, and economic burdens associated with severe mental illness prompt researchers to search for new therapies and also to develop accurate methods of differential diagnosis, which should be ultimately based on objective, biological markers (Pantelis et al., 2009). The development of new neuroimaging techniques enables researchers to identify neural circuits that underline the human brain integration system. Various neuropsychiatric conditions are correlated with changes in brain communication patterns and pointed as potentially useful biomarkers for clinical applications (Sporns et al., 2005). In accordance with the results of earlier studies focused on brain synchronization, SZ is seen unequivocally as a disconnectivity disorder characterized by abnormal functional and structural connectivity of the brain (Friston and Frith, 1995). Application of diffusion tensor imaging (DTI) methods, such as magnetic resonance imaging (MRI) technique, into SZ research, showed disconnection and multiple microstructural aberrances of brain white matter fibers (Zalesky et al., 2011; Klauser et al., 2017). Studies based on electroencephalography (EEG) and functional MRI (fMRI) also revealed abnormalities in the functional connectivity of the brain, which were also correlated with the clinical picture of the SZ (Skudlarski et al., 2010; Uhlhaas, 2013; Krukow et al., 2018). Nevertheless, to understand the systemic level of the brain organization and to explain neurophysiological processes such as disconnectivity syndrome in the SZ, researchers started to analyze the brain as a complex network (van den Heuvel and Sporns, 2013). The neural network is understood as a system of spatial (anatomical) and temporal (synchronous firing of neuronal assemblies) dimensions, involving different brain regions interconnected with each other (Zalesky et al., 2010). However, to analyze the state of the functional and structural connections from the viewpoint of the entire brain, an infinite number of potential anatomical and functional interactions between a given set of neural regions makes such an analysis a challenge almost impossible to obtain. Therefore, a graph theorem has been introduced to solve this problem and to test the complex whole-brain networks in their global dimension (Van Den Heuvel and Fornito, 2014). Previous studies investigating the neural brain networks in SZ showed significantly changed network organization as indicated by graph-analytical measures of global, short communication paths (Yan et al., 2015), local organization (Alexander-Bloch et al., 2010), and small-worldness (balance between local segregation and global integration) (Shim et al., 2014). Aberrant functional networks in the SZ were also linked with cognitive impairments (Sheffield et al., 2015; Krukow et al., 2020) and the duration of the illness (Jonak et al., 2019).

Previous studies considered the problem of automated classification of altered brain activity in SZ based on the EEG or fMRI data. Among traditional classifiers, methods such as support vector machine (SVM; Shim et al., 2016; Liu et al., 2017; Huang et al., 2018), adaptive boosting (Sabeti et al., 2011), kernel discriminant analysis (KDA; Zhu et al., 2018), or nearest neighbor algorithm (Parvinnia et al., 2014) are used. Some of these studies (Sabeti et al., 2011; Parvinnia et al., 2014) applied time-frequency features obtained from single EEG channels, which is a limited capacity approach as it does not consider interactions between channels understood as a network. Other authors applied a convolutional neural network (Phang et al., 2019) and deep neural networks (DNNs; Plis et al., 2014; Guo et al., 2017). In addition, manifold learning for aggregation was considered in works by Shen et al. (2010); Anderson and Cohen (2013), and Gallos et al. (2021a,b). The idea of applying fuzzy classification into SZ-based data is a relatively new concept, as there are only a few papers on this subject (Sabeti et al., 2007; Silvana et al., 2018). One of the answers to the problems related to the application of single classifiers in the processes of automated disease diagnosis may be using various aggregation models. Aggregation can be carried out at the stage of data analysis in the form of information fusion and the stage of analysis of classification results. Despite some shortcomings such as extending the duration of the diagnosis process or the need to implement additional algorithms, the undoubted advantage of this approach is the increase in the effectiveness of classification, which, combined with the field of application critical to human health, is of key importance. Common examples of aggregation operators are voting, maximum, minimum, and median functions. The methods based on triangular norms (Klement et al., 2000) or ordered weighted averaging operators (OWA; Yager and Kacprzyk, 2012) are somewhat more complex. Various general approaches to the aggregation of classifiers were already presented (e.g., in publications of Alsina et al., 2006; Beliakov et al., 2007; Grabisch et al., 2009; Calvo et al., 2012; Gągolewski, 2015; Dolecki et al., 2016; Baczyński et al., 2017). Recently, one of the dominant techniques is using the Choquet integral or its generalizations or extensions (Kwak and Pedrycz, 2004, 2005; Karczmarek et al., 2014, 2017a,b, 2018, 2019b; Anderson et al., 2018; Rutkowska et al., 2020). In particular, recent studies on the so-called pre-aggregation functions offer hope for the development of this approach (Lucca et al., 2015, 2016, 2017; Bustince et al., 2016; Dimuro et al., 2017; Dias et al., 2018). They are particularly used in computer image analysis and its subdiscipline of facial recognition (Karczmarek, 2018; Karczmarek et al., 2019a). Detailed theoretical and practical

analyses of the approach based on pre-aggregation functions, i.e., slightly weakening the classical aggregation (Beliakov et al., 2007) conditions, are still ongoing. Nevertheless, the weakening of these conditions does not have a negative impact on the classification results, which is confirmed by the experimental outcomes from the above-mentioned studies. A survey of the generalizations of Choquet integral can be found in Dimuro et al. (2020).

The problem undertaken in this study was related to the effective automatic distinction between patients diagnosed with SZ and healthy subjects based on EEG-based features of neuronal network organization. The main goal of this study was to find the appropriate operator aggregating the neurophysiological outcomes and categorizing them as patients diagnosed with SZ or healthy controls (HC), i.e., increasing the effectiveness of the classification. For this purpose, various generalizations of the Choquet integral were tested and a set of over a thousand aggregating functions not related to the Choquet integral was verified. In the section on numerical experiments, we indicate the classes of functions and their detailed parameters that work best in terms of the identification of SZ. The dataset applied in this study included data gathered from 40 subjects, i.e., 20 schizophrenic patients and 20 HCs. The Granger causality (GC; Granger, 1969) concept had been applied to particular EEG bands to achieve functional brain connectivity measures. The collected measurements were analyzed using a minimum spanning tree (MST; Stam et al., 2014). The global MST parameters obtained in the analysis were chosen as features in the classified dataset. Applying the MST algorithms enabled grasping the backbone structure of the brain network with only the strongest connections included (González et al., 2016; Van Dellen et al., 2016). Using MST ensured that the link between nodes was not based on an arbitrarily set connectivity strength threshold, which allowed avoiding the bias in network density computations (Tewarie et al., 2015). In general, the MST parameters were chosen because this method lacks some theoretical and mathematical problems incorporated to more typical network organization indicators based on the small-worldness approach and, above all, because the authors wanted to refer to the concept of SZ as a disconnection disease in which pathology of neuronal integration is not isolated only to selected regions or type of synchronizations but has a global dimension. The MST enables the characterization of the global, whole-brain network.

MATERIALS AND METHODS

Participants

Twenty patients, who met the *DSM-5* (*Structured Clinical Interview for DSM-5*) criteria for SZ, were involved from the Department of Psychiatry, Psychotherapy and Early Intervention, Medical University of Lublin. Additionally, the other criteria were as follows: age over 18 years; minimum 10 years of regular education; not more than 5 psychiatric hospitalizations associated with exacerbation of psychosis; no markers of structural brain abnormalities visualized on MRI, indicative of surviving craniocerebral trauma or neurovascular episodes; and lack of serious somatic diseases needing intense pharmacotherapy that would impact the EEG recordings. During testing, all patients were on stable doses of atypical antipsychotics. Using anticholinergic agents, benzodiazepines, and mood stabilizers up to 3 months before the assessment was an exclusionary factor for all participants. The patients participated in the study during the last week of psychiatric hospitalization, after obtaining a significant clinical and functional improvement, being fully able to give consent and undergo EEG examination. The control group consisted of HC, demographically matched to the clinical group, who were chosen from the local community. Additionally, HC had no history of psychiatric diagnoses, per the *Structured Clinical Interview for DSM-5*, brain disease, or neurological injury as well as no family history of psychosis. All patients consented to the study in accordance with the protocol approved by the Bioethical Commission of the Medical University of Lublin. The Commission also validated the methods used in the study. Demographical and clinical data are presented in **Table 1**. The groups did not differ significantly in terms of age (SZ = 34.41, *SD* = 8.41; HC = 31.63, *SD* = 6.42), number of the years of education (SZ = 12.43, *SD* = 2.94; HC = 14.87, *SD* = 1.68), and gender (SZ = 50% of men; HC = 50% of men). In the SZ group, the duration of illness lasted for about 12 years.

Data Acquisition

Using a 21-scalp position, electro-cap electroencephalograph (Electro-Cap International Inc., OH, United States) and Ag/AgCl disk electrodes, in 10 min of resting-state, EEG data were recorded for each participant. Electrodes were distributed according to the 10–20 International system (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, A1, A2, F7, F8, T3, T4, T5, T6, Fz, Pz, and, Cz). Subjects were seated with eyes closed and restricted head movement. The electrode impedances were kept below 5 k, and the data were filtered from 0.5 to 70 Hz (with active notch filter set at 50 Hz) when the sampling rate was 512 Hz. The data were exported into ASCII format after recording. Post-processing procedures were carried out in the EEGLAB program, which is a MATLAB toolbox. First, the signal was filtered using the bandpass filter at 0.5–45 Hz (second-order Butterworth filter). Second, the reference was changed offline into the averaged. Next, from the processed signal, 25 epochs lasting for 8 s (4,096 samples) without artifacts were extracted for each patient by a clinical neurophysiologist. Last, EEG signals were divided into six

TABLE 1 | Demographic and clinical data of research groups.

	SZ (n = 20) M (SD)	*HC (n = 20) M (SD)*	*z value or* χ^2	*p*
Age (years)	32.41 (8.41)	31.63 (6.42)	0.16	0.91
Education (years)	12.43 (2.94)	14.87 (1.68)	−1.12	0.45
Sex (% male)	50	50	0	1
Duration of illness (years)	12.1 (9.43)			
Number of hospitalizations	2.25 (2.65)			
Risperidone equivalents	4.66 (1.76)			

frequency bands using finite impulse response filters: the delta (0.5–4 Hz), theta (4–8 Hz), low alpha (8–10 Hz), high alpha (10–12 Hz), beta (13–30 Hz), and gamma (30–45 Hz).

Data Processing

Several steps were involved in the data processing procedure. Feature extraction was associated with the calculation of functional brain connectivity (FC) measures separately for particular electrodes and each EEG frequency band. The FC was calculated to indicate the statistical dependence between the spatially distributed neurophysiological time series such as EEG signals stemming from separate units of a nervous system (Cheng et al., 2015). There were several metrics used in assessing FC strength, such as classical measures (e.g., Pearson's correlation coefficient, cross-correlation function, or coherence), phase synchronization indexes (e.g., phase lag index or phase-locking value), and GC measures. We chose the classical linear GC. The idea of GC (Granger, 1969) was based on the assumption that having two simultaneously determined signals (X and Y), the signal X could be better explained using information from the signal Y than using only information from the signal X. In such a situation, signal Y could be specified as "causal" to signal X. The GC measure is widely applied as a statistical tool to detect the influence of particular system components (Nolte et al., 2010). For the GC calculation, we used the Matlab MVGC toolkit (Sackler Centre for Consciousness Science, University of Sussex, Brighton, United Kingdom; Barnett and Seth, 2014), which was based on advanced VAR (vector autoregressive) model theory. To optimize auto-covariance delays, the Akaike information criterion was used to estimate the optimal model order. MVGC algorithms were used to convert EEG signals into auto-covariance data. The observed auto-covariance sequences were then subjected to paired spectral GC. In the case study, the calculated GC measures were used in the next feature extraction procedure step consisting of deriving MST. The MST was built based on Kruskal's algorithm. First, the weight of all edges was sorted, and then the stronger edges were connected, i.e., those with the highest connectivity values, eliminating those that form loops. These stages were repeated as many times as necessary until the final tree had 19 nodes and 18 edges, which corresponded to the total number of electrodes used in our EEG recording. The MST metrics were generated separately for each frequency band.

Global MST parameters were used as the features for the classification procedure, which included the following:

- maximal degree, a maximal degree in the MST tree,
- maximal BC, maximal betweenness centrality in the MST tree,
- leaf fraction (Lf), the ratio between leaf vertex number (further denoted as L) and the total vertex number,
- diameter (d), the longest distance between any two vertices in the MST tree,
- hierarchy (Th), the measure describing the optimality of the tree topology.

After the feature extraction procedure, the classification was done using classical classifiers.

A General Processing Scheme

The classification procedure was performed to assign observations into one of two classes: schizophrenic patient and HC. Several classical classification methods were used: cubic SVM, linear SVM, decision tree, logistic regression, multilayer perceptron (MLP), random forest, and k-nearest neighbors (kNN). Data were split into training and testing datasets based on the 80:20 ratio.

The results of this classification process, in the form of the probabilities of belonging to considered classes, were taken as the inputs to the aggregation functions. The next step of the analysis procedure was to generate fuzzy measure density values to apply aggregation operators. The aggregation operators allow combining the predictions of multiple classifiers to further improve the results. The fuzzy measures can be interpreted as the degree of trust (weights or level of importance) to predictions of the individual classifier.

In general, there are several methods of fuzzy measure generation, such as expert assumption, optimization, and the heuristic one. In our study, the cross validation-based heuristic one was applied. N-fold cross-validation was run on the training set to obtain a density measure for a classifier.

It is a well-known fact that the results of different classifications can be aggregated. This situation can be easily illustrated by the example of various sports competitions held in the form of a Grand Prix cycle, where the points are added together to determine the final winner. EEG signal-based features can be similarly aggregated. In general, the results obtained using different kinds of classifiers can be added, averaged, or, generally speaking, transformed with the help of different kinds of aggregation operators. Typical operators are median, minimum, and maximum functions, etc. The general scheme of classification using aggregation methods is presented in **Figure 1**. It is worth noting that the values that are input to the aggregation operator can change the distances between the training and testing representation of an EEG signal in the case of *k*-nearest neighbor-based classifiers, likelihoods of belonging to a class in the case of neural network-like methods, etc. Here, it is worth stressing that the weights presented in this diagram can be obtained from experts but also based on the quality of classification of individual classifiers (e.g., their accuracy measures).

Aggregation of Classifiers Using the Choquet Integral and Its Extensions

One of the best known and most efficient classifiers is the Choquet integral. Hence, let us recall the main properties of the fuzzy measure, Choquet integral, and its generalizations. Let us denote a set as X. Then $P(X) = 2^X$ is a family of all its subsets. 2^X is a σ-algebra; i.e., the empty set belongs to it, the complement of a set belonging to σ-algebra belongs to it, and the sum of countable many sets from the σ-algebra also belongs to it. Generally speaking, in the context of classification tasks, the elements of the set X are the individual classifiers (methods, parts of the images under consideration, etc.). In the context of this specific study, they are particular classifiers, see the experimental section for details of the methods. These classifiers are denoted as $x_1, x_n, \quad n \quad 1$. Now, one can define (Sugeno, 1974) a fuzzy

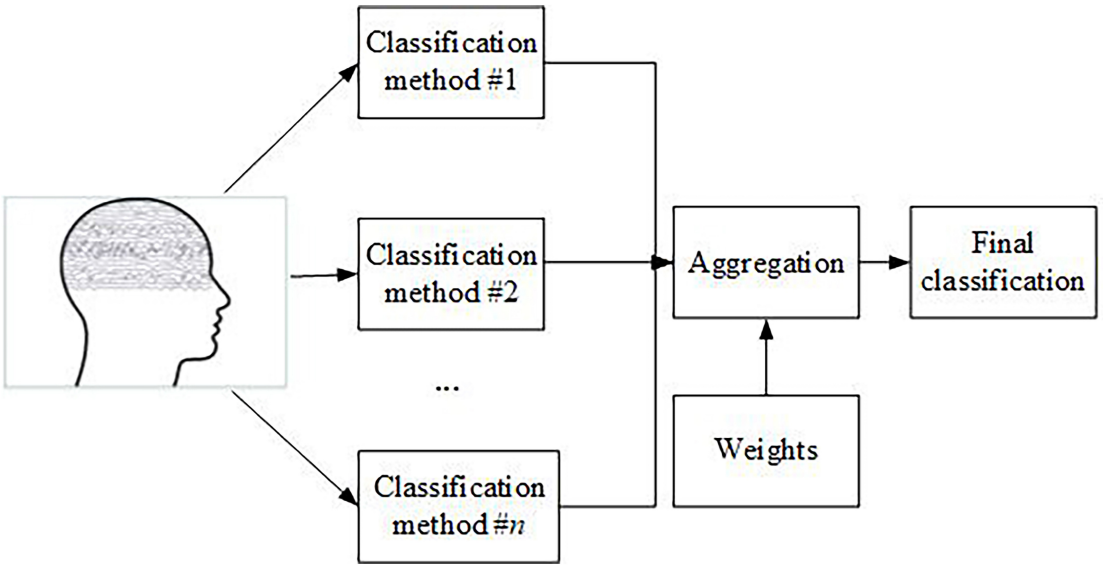


FIGURE 1 | A general classification scheme based on an aggregation process.

measure as a set function $g : P(X) \to \mathbb{R}$ satisfying the following conditions:

$$g(\emptyset) = 0 \text{ and } g(X) = 1 \tag{1}$$

$$g(U) \leq g(W), U \subset W, U, W \in P(X) \tag{2}$$

$$\lim_{n\to\infty} g(U_n) = g\left(\lim_{(n\to\infty)} U_n\right) \tag{3}$$

Here, $\{U_n\}$, $n = 1, 2, \ldots$ means increasing set sequence. Recall that Sugeno λ-fuzzy measure realizes the above conditions and

$$g(U \cup W) = g(U) + g(W) + \lambda g(U) g(W) \tag{4}$$

with $\lambda > -1$. Here, U and W are not overlapping. In addition, we have

$$g(U_{i+1}) = g(U_i) + g_{i+1} + \lambda g(U_i) \tag{5}$$

for $U_i = \{x_1, \ldots, x_i\}$, $U_{i+1} = \{x_1, \ldots, x_{i+1}\}$. The following notation is used commonly: $g_i = g(\{x_i\})$, $i = 1, \ldots, n$. Now, let us introduce a function $h(x)$ and let the series $h(x_i)$, $i = 1, \ldots, n$, be ordered non-increasingly and let $h(x_{n+1}) = 0$. In the context of this study, the function $h(\cdot)$ represents a value of classifier describing the probability of belonging to a specific class. Next, the U_i set is, in fact, only an abstract object. The real importance has the value of $g(U_i)$ appearing in (5) which can be easily found recursively starting from the values of g_i. The value g_i represents a significance (or importance) of a particular classifier x_i. Its value can be commonly defined twofold: (1) based on the opinions of experts and (2) based on initial tests. In this study, we applied the second method. Finally, n is a number of classifiers. The last parameter to be found is λ, which can be obtained from the following equation:

$$1 + \lambda = \prod_{i=1}^{n} (1 + \lambda g_i), \; g_i = g(\{x_i\}) \tag{6}$$

see Sugeno (1974).

For such assumptions, the Choquet integral is defined as

$$C = \sum_{i=1}^{n} \left(h(x_i) - h(x_{i+1}) g(U_i)\right) \tag{7}$$

From this function, many generalizations and extensions can be delivered as follows:

$$C_M = \sum_{i=1}^{n} M\left(h(x_i) - h(x_{i+1}), g(U_i)\right) \tag{8}$$

and for any t-norm M (·,·), see Lucca et al. (2014),

$$C_{FM} = \min\left(\sum_{i=1}^{n} M\left(h(x_i) - h(x_{i+1}), g(U_i)\right), 1\right) \tag{9}$$

(Lucca et al., 2014, 2015),

$$C_{CM} = \sum_{i=1}^{n} \left(M\left(h(x_i), g(M_i)\right) - M\left(h(x_{i+1}), g(U_i)\right)\right) \tag{10}$$

see (Lucca et al., 2017), C_{Min} (Lucca et al., 2015), where the role of the function M is played by the minimum, or C_O [see (Lucca et al., 2016)] with a so-called overlap function under the integral

sign. Newer functions were proposed in Karczmarek (2018) and Karczmarek et al. (2019a). They are

$$C_{MC} = \sum_{i=1}^{n} \left(M\left(h\left(x_i\right), g\left(U_i\right)\right) - M\left(h\left(x_{i+1}\right), g\left(U_i\right)\right) + M\left(h\left(x_i\right) - M\left(x_{i+1}\right), g\left(U_i\right)\right)\right) \quad (11)$$

$$C_{MMin} = \sum_{i=1}^{n} M\left(\min\left(h\left(x_i\right), g\left(U_i\right)\right) - \min\left(h\left(x_{i+1}\right), g\left(U_i\right)\right), g\left(U_i\right)\right) \quad (12)$$

$$C_{MMin2} = \sum_{i=1}^{n} M\left(\min\left(h\left(x_i\right), g\left(U_i\right)\right), \min\left(h\left(x_{i+1}\right), g\left(U_i\right)\right)\right) \quad (13)$$

$$C_{M\mathrm{in}\mathrm{M}} = \sum_{i=1}^{n} \min\left(M\left(\left(h\left(x_i\right), g\left(U_i\right)\right)\right), M\left(h\left(x_{i+1}\right), g\left(U_i\right)\right)\right) \quad (14)$$

and the integrals inspired by some numerical analysis formulae such as

$$C_{D1} = \sum_{i=1}^{n} M\left(h\left(x_{i-1}\right) - h\left(x_{i+1}\right), g\left(U_i\right)\right) \quad (15)$$

$$C_{D2} = \sum_{i=1}^{n} M\left(h\left(x_{i-1}\right) + h\left(x_{i+1}\right) - h\left(x_i\right), g\left(U_i\right)\right) \quad (16)$$

and

$$C_{D3} = \sum_{i=1}^{n} M\left(\frac{h\left(x_{i-1}\right) + h\left(x_{i+1}\right)}{h\left(x_i\right)}, g\left(U_i\right)\right) \quad (17)$$

It is worth noting that $M(\cdot,\cdot)$ can be any triangular norm which, as an intersection or conjunction operator in many application areas, is a counterpart to a classic product operator appearing in the original Choquet integral.

RESULTS

Individual Classifiers

In this study, we described particular classifiers that were considered in the series of numerical experiments and determined their accuracy. We applied the following classical machine learning models: SVM with linear and cubic kernels, logistic regression, kNN, decision tree, random forest, and MLP. The classical machine learning models were used due to a low number of observations available for training and testing. In order to obtain the fuzzy density that is necessary for aggregation, the following approach can be adapted. According to the holdout validation procedure, the data were split into training and validation subsets, where 20% of the dataset was used for validation. The fuzzy density was calculated as a mean accuracy measure obtained in the process of a fivefold cross-validation run using the training data. The resulting classification quality of separate models was tested on the validation subset after fitting the models on the complete training set. The classification accuracy values obtained with separate classifiers are presented in **Figure 2**.

Aggregation of Classifiers

The experimental results of the aggregation scheme used for the classifiers discussed in the previous section, namely decision tree, k-nearest neighbor, quadratic SVM, cubic SVM, linear SVM, logistic regression, random forest, and MLP, are discussed. The accuracies of the individual classifiers obtained in the initial series of experiments are the input to establish the fuzzy measure densities g_i. The values of the function h are the results of the classification of testing elements being the probabilities of belonging to the two classes, namely, healthy and SZ patients. The validation procedure described in the previous section was repeated 200 times. After each run, a value of fuzzy density and 8 probability vectors (20% out of 40 observations) were obtained per classification model. The details of aggregation algorithm implementation required a single estimation of classification accuracy per model and aggregation method. Hence, all 1,600 (200×8) classification results were analyzed. It is worth noting that the models were fitted and the fuzzy densities were obtained independently in every separate run of the experiment. In the series of experiments, we have evaluated 25 classes (families) of popular and commonly considered in the literature triangular norms (Alsina et al., 2006, page 72). The monograph can be treated as a compendium of the t-norms to be applied in more advanced aggregation operators. In this particular approach, the t-norms serve as the integer functions M with parameters -10, -9.9, ..., 0, ..., 9.9, 10, but only if the parameter is in the range allowed for the t-norm. Such a choice of the parameter range seems to be optimal and emphasizes the most important properties of each of the t-norm classes. The maximal accuracy was obtained for the classifier C_{D2} for triangular norm from the family no. 8, namely

$$M\left(x, y\right) = \frac{\max\left(\alpha^2 xy - (1-x)\left(1-y\right), 0\right)}{\alpha^2 - (\alpha-1)^2\left(1-x\right)\left(1-y\right)}, \alpha > 0 \quad (18)$$

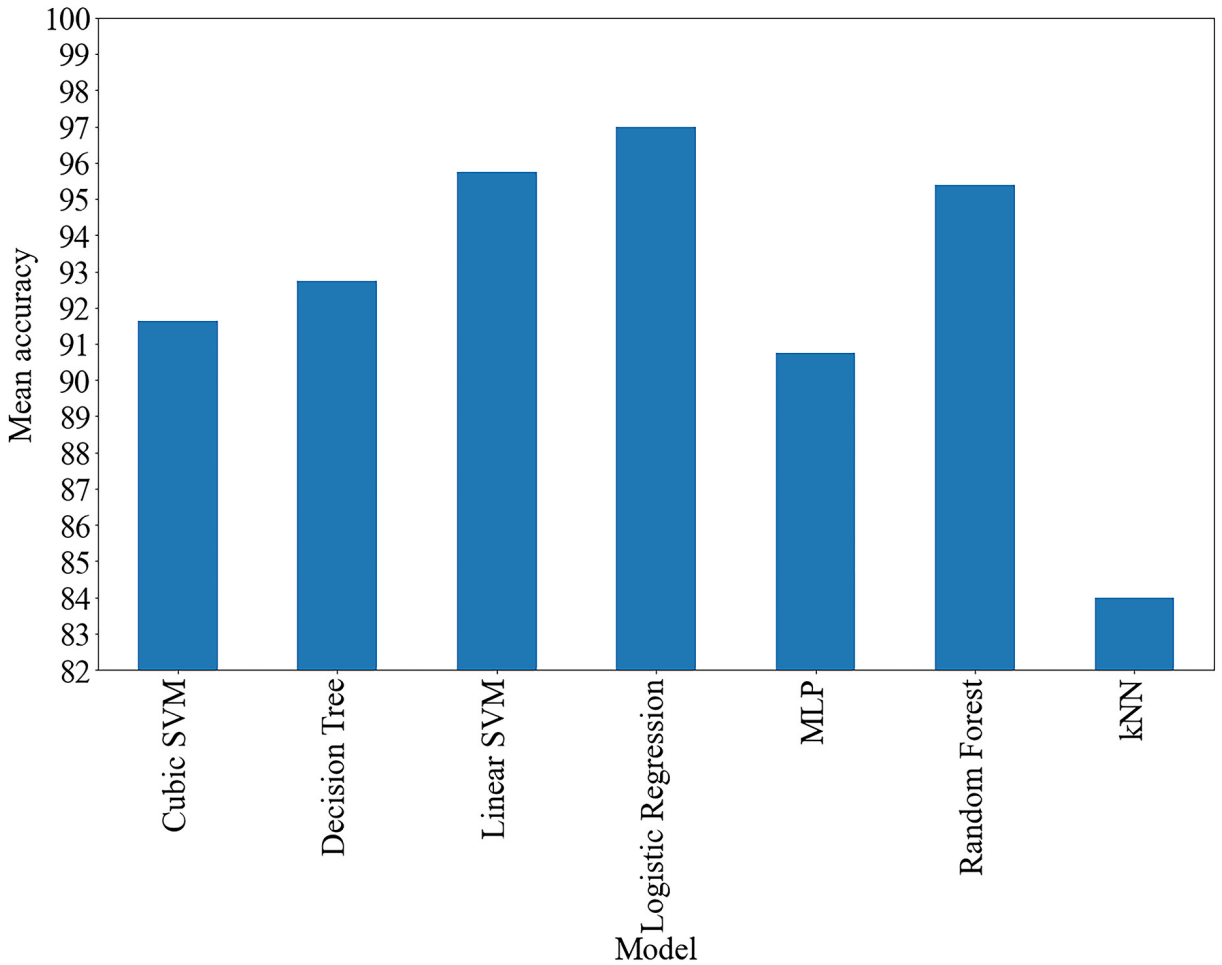


FIGURE 2 | Average of accuracies of separate classifiers.

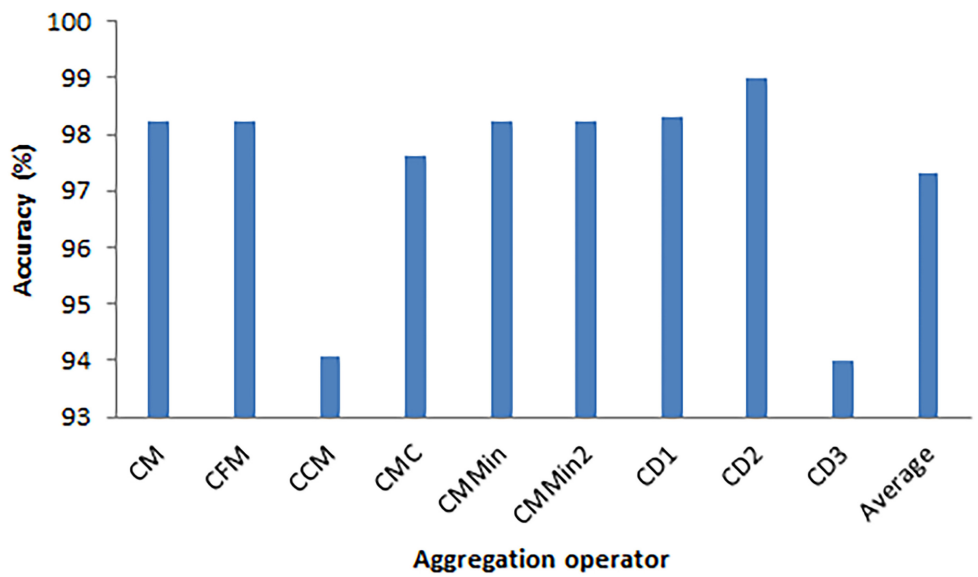


FIGURE 3 | The accuracies were obtained with the function (18) and aggregation operator C_{D2}.

TABLE 2 | Triangular norms and their parameters for the results.

Aggregation function	Number of *t*-norm family	α
C_M	12	0.2
C_{FM}	12	0.2
C_{D2}	2	0.5
C_{D2}	8	0.1, 0.2
C_{D2}	11	2.1, 2.4
C_{D2}	13	1.2, 1.3
C_{D2}	14	2.1, 2.4
C_{D2}	25	4.4, 5, 5.1
Average	8	0.1

TABLE 3 | The best choices of *t*-norms for various generalizations of the Choquet integral.

Aggregation function	Number of *t*-norm family
C_M	12
C_{FM}	12
C_{CM}	11, 14
C_{MC}	11, 14
C_{MMin}	4, 12
C_{MMin2}	4, 12
C_{D1}	4, 13
C_{D2}	8
C_{D3}	4

In this case, the best option to choose is α = 0.1. The plot illustrating the recognition rate for the combination of C_{D2} and the function (18) is given in **Figure 3**. It is obvious that satisfying accuracy can be obtained only for relatively small values of the parameter α.

It is important to understand the process of calculation of the value of the aggregation function. Let us consider, for example, the process of C_{D2} finding. Here, the individual classifiers x_i, $i = 1, \ldots, n = 5$ (five classifiers are discussed in this experiment) should be analyzed, and for their significance measures (simply, weights) $g_i = g(\{x_i\})$, see **Figure 2**. Next, the parameter λ appearing in (6) is calculated. Based on the value of λ, the values of $g(U_i)$ appearing in Eq. (5) are found recursively. Next, using the values of $h(x_{i-1})$, which are the likelihoods of belongings of a given probe to a specific class, the final sum (16) can be obtained, taking into account that $M(\cdot, \cdot)$ is any *t*-norm, in particular a function given by (19).

Very good results were obtained also for the aggregation functions C_M and very similar C_{FM}. Maximal yielded values were 98.81% for the function number 12 serving as integer function and α = 0.2. The formula of the *t*-norm is as follows:

$$M(x, y) = \max\left(1-\left((1-x)^{\alpha} + (1-y)^{\alpha}\right)^{\frac{1}{\alpha}}, 0\right) \alpha > 0 \quad (19)$$

Table 2 supplements the above discussion by showing for which triangular norms and their parameters the classification rate exceeding 98.81% was reached.

It is worth stressing that the best average result among the operators C_M, C_{FM}, C_{CM}, C_{MC}, C_{MMin}, C_{MMin2}, C_{D1}, C_{D2}, and C_{D3} was also obtained for the triangular norm no. 8 given by the formula (18) and its parameter α = 0.1. The plot presenting the values of the combination of all the aggregation functions with this *t*-norm and parameter is presented in **Figure 4**. It is obvious that the function (18) works well with almost all of the aggregation operators except C_{CM} and C_{D3}.

As a supplement to the results, it is worth noting that **Table 3** presents the information for which the best results of the *t*-norms

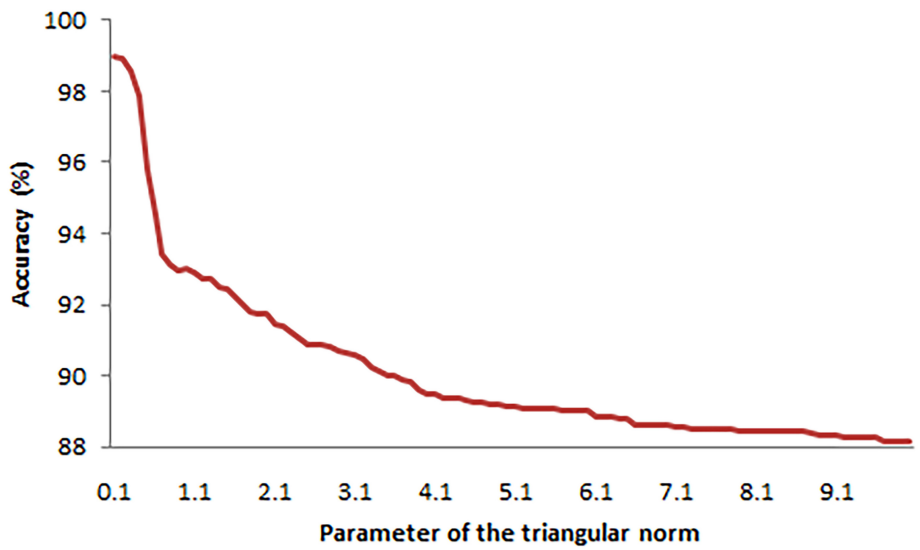


FIGURE 4 | Averages of accuracies achieved with top aggregation functions.

were obtained by aggregation operators. It can help match t-norms and generalizations of the Choquet integral by experts conducting similar research. For instance, function no. 4, see Alsina et al. (2006), works well when it is combined with a few Choquet integral-based operators.

Finally, what should be emphasized, we have conducted the same tests for over 1,000 functions that are not Choquet-like integrals. The functions that were used in this competition were selected based on the studies by Alsina et al. (2006), Beliakov et al. (2007); Grabisch et al. (2009), and Calvo et al. (2012). The best results were obtained for the ordinary weighted averaging operator at the level of 98.81%.

$$OWA(x_1, \ldots, x_n) = \sum_{j=1}^{n} \omega_j y_j \tag{20}$$

where y_j is the j-th largest of the x_i, and the weights are $\omega_1 = 1$, $\omega_2 = 1 - \frac{1}{n}, \ldots, \omega_n = \frac{1}{n}$ with or without normalization to their sum. Despite it being a very good result, it is obvious that it is hard to find the function giving the results more satisfying than Choquet integral-based operators. Moreover, to find the proper form of OWA, similar to the Choquet integral case, a proper heuristic can be used.

CONCLUSION AND FUTURE WORK

In this study, we have indicated the most appropriate operator aggregating the results of binary classification of patients to efficiently distinguish individuals with SZ and healthy subjects using a set of neural network organization features extracted from EEG-based functional connectivity measures. A series of both types of functions, generalizations of the Choquet integral and other aggregating functions, have been verified to determine the classes of functions and their parameters, which are the most effective in the classification of SZ. As an input to the main analysis, the results of classification were performed with classical methods such as decision tree, k-nearest neighbor, quadratic SVM, cubic SVM, linear SVM, logistic regression, random forest, and MLP were applied. The original results obtained in the study of classical methods classification reached 97% for logistic regression. Although the initially obtained results were high, we decided to verify if there is the possibility to reach even higher results using the fuzzy-based classifier.

The results prove that applying various classification models in combination with aggregation functions enable further improvement of classification results. This approach allows us to take advantage of the additional knowledge cumulated in the parameters of the trained models.

Detailed results show that several aggregation functions enabled to give promising results [presented in the study as Eqs (9–13) and (15, 16)], which increase the classification result by more than 1%. Among numerous functions evaluated and implemented in the thorough comparison, the best accuracy was reached for the aggregating integral C_{D2} with triangular norm appearing under the integral sign given by the formula (19). Very good results (classification accuracy higher than 98.8%) were reached also for aggregation functions CM and CFM. It is worth noting that the obtained results occurred to be better than the original accuracy reached with classical methods by 1.81%. Although the obtained improvement is not very high (less than 2%), the overall increase in classification accuracy from 97% (for the best classical classifier) to as high as almost 99% (for the properly selected pre-aggregation operators) is relevant. Nevertheless, we have not done double cross-validation analysis, so this limitation can influence classification accuracy, hence the classification rate could be slightly overestimated. Results show the usefulness of this method especially if the role of aggregation function is an extended version of the Choquet integral. In contrast, the application of aggregation functions could give a relatively better improvement in case of weaker initial individual classification results. In future, we have planned to extend the analysis to consider more phases and stages of SZ. Moreover, we are interested in the application of other classes of aggregation operators and the determination of their weights (significance in the process of aggregation) based on the opinions of medical experts. More theoretically, it is still an interesting as well as difficult task to find the optimal parameters of the integral operators only based on their results according to various classification tasks with no relation to the accuracies or expert opinions. Finally, an application of aggregation techniques in other medical pattern recognition or classification problems will be worth analyzing.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Bioethics Committee at the Medical University of Lublin. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MP-W and MK: contributed in methodology, software, validation, writing—original draft preparation, and visualization. PKa: contributed in conceptualization methodology, software, writing—original draft preparation, and visualization. MT: contributed in methodology, software, writing—original draft preparation, and visualization. PKr: contributed in methodology, validation, and writing—original draft preparation. KJ: contributed in methodology and writing—original draft preparation. All authors contributed to the article and approved the submitted version.

REFERENCES

Alexander-Bloch, A. F., Gogtay, N., Meunier, D., Birn, R., Clasen, L., Lalonde, F., et al. (2010). Disrupted modularity and local connectivity of brain functional networks in childhood-onset schizophrenia. *Front. Syst. Neurosci.* 4:147. doi: 10.3389/fnsys.2010.00147

Alsina, C., Schweizer, B., and Frank, M. J. (2006). *Associative Functions: Triangular Norms and Copulas*. Singapore: World Scientific.

Anderson, A., and Cohen, M. S. (2013). Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: an fMRI classification tutorial. *Front. Hum. Neurosci.* 7:520. doi: 10.3389/fnhum.2013.00520

Anderson, D. T., Scott, G. J., Islam, M. A., Murray, B., and Marcum, R. (2018). "Fuzzy choquet integration of deep convolutional neural networks for remote sensing," in *Computational Intelligence for Pattern Recognition*. (Cham: Springer), 1–28. doi: 10.1007/978-3-319-89629-8_1

Baczyński, M., Bustince, H., and Mesiar, R. (2017). Aggregation functions: theory and applications, part I. *Fuzzy Sets Syst.* 324, 1–2.

Beliakov, G., Pradera, A., and Calvo, T. (2007). *Aggregation Functions: A Guide for Practitioners*, Vol. 221. Heidelberg: Springer.

Barnett, L., and Seth, A. K. (2014). The MVGC multivariate Granger causality toolbox: a new approach to Granger-causal inference. *J. Neurosci. Methods* 223, 50–68.

Bustince, H., Sanz, J. A., Lucca, G., Dimuro, G. P., Bedregal, B., Mesiar, R., et al. (2016). "Pre-aggregation functions: definition, properties and construction methods," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. (Piscataway, NJ: IEEE), 294–300.

Calvo, T., Mayor, G., and Mesiar, R. (eds) (2012). *Aggregation Operators: New Trends and Applications*, Vol. 97. Heidelberg: Physica.

Cheng, W., Palaniyappan, L., Li, M., Kendrick, K. M., Zhang, J., Luo, Q., et al. (2015). Voxel-based, brain-wide association study of aberrant functional connectivity in schizophrenia implicates thalamocortical circuitry. *NPJ Schizophr.* 1, 1–8.

Dias, C. A., Bueno, J. C., Borges, E. N., Botelho, S. S., Dimuro, G. P., Lucca, G., et al. (2018). "Using the Choquet integral in the pooling layer in deep learning networks," in *North American Fuzzy Information Processing Society Annual Conference*. (Cham: Springer), 144–154. doi: 10.1007/978-3-319-95312-0_13

Dimuro, G. P., Fernández, J., Bedregal, B., Mesiar, R., Sanz, J. A., Lucca, G., et al. (2020). The state-of-art of the generalizations of the Choquet integral: from aggregation and pre-aggregation to ordered directionally monotone functions. *Inform. Fusion* 57, 27–43. doi: 10.1016/j.inffus.2019.10.005

Dimuro, G. P., Lucca, G., Sanz, J. A., Bustince, H., and Bedregal, B. (2017). "CMin-Integral: a Choquet-like aggregation function based on the minimum t-norm for applications to fuzzy rule-based classification systems," in *International Summer School on Aggregation Operators*. (Cham: Springer), 83–95. doi: 10.1007/978-3-319-59306-7_9

Dolecki, M., Karczmarek, P., Kiersztyn, A., and Pedrycz, W. (2016). "Utility functions as aggregation functions in face recognition," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. (Piscataway, NJ: IEEE), 1–6.

Friston, K. J., and Frith, C. D. (1995). Schizophrenia: a disconnection syndrome. *Clin. Neurosci.* 3, 89–97.

Gągolewski, M. (2015). *Data Fusion: Theory, Methods, and Applications*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Gallos, I. K., Galaris, E., and Siettos, C. I. (2021a). Construction of embedded fMRI resting-state functional connectivity networks using manifold learning. *Cogn. Neurodyn.* 15, 585–608. doi: 10.1007/s11571-020-09645-y

Gallos, I. K., Gkiatis, K., Matsopoulos, G. K., and Siettos, C. (2021b). ISOMAP and machine learning algorithms for the construction of embedded functional connectivity networks of anatomically separated brain regions from resting state fMRI data of patients with Schizophrenia. *AIMS Neurosci.* 8, 295–321. doi: 10.3934/Neuroscience.2021016

González, G. F., Van der Molen, M. J. W., Žarić, G., Bonte, M., Tijms, J., Blomert, L., et al. (2016). Graph analysis of EEG resting state functional networks in dyslexic readers. *Clin. Neurophysiol.* 127, 3165–3175. doi: 10.1016/j.clinph.2016.06.023

Grabisch, M., Marichal, J. L., Mesiar, R., and Pap, E. (2009). *Aggregation Functions (No. 127)*. Cambridge: Cambridge University Press.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–438. doi: 10.2307/1912791

Green, M. F., Horan, W. P., and Lee, J. (2019). Nonsocial and social cognition in schizophrenia: current evidence and future directions. *World Psychiatry* 18, 146–161. doi: 10.1002/wps.20624

Guo, X., Dominick, K. C., Minai, A. A., Li, H., Erickson, C. A., and Lu, L. J. (2017). Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Front. Neurosci.* 11:460. doi: 10.3389/fnins.2017.00460

Heilbronner, U., Samara, M., Leucht, S., Falkai, P., and Schulze, T. G. (2016). The longitudinal course of schizophrenia across the lifespan: clinical, cognitive, and neurobiological aspects. *Harvard Rev. Psychiatry* 24, 118–128. doi: 10.1097/HRP.0000000000000092

Huang, J., Zhu, Q., Hao, X., Shi, X., Gao, S., Xu, X., et al. (2018). Identifying resting-state multifrequency biomarkers via tree-guided group sparse learning for schizophrenia classification. *IEEE J. Biomed. Health Inform.* 23, 342–350. doi: 10.1109/JBHI.2018.2796588

Jonak, K., Krukow, P., Jonak, K. E., Grochowski, C., and Karakuła-Juchnowicz, H. (2019). Quantitative and qualitative comparison of EEG-based neural network organization in two schizophrenia groups differing in the duration of illness and disease burden: graph analysis with application of the minimum spanning tree. *Clin. EEG Neurosci.* 50, 231–241. doi: 10.1177/1550059418807372

Karczmarek, P. (2018). *Selected Problems of Face Recognition and Decision-Making Theory*. Lublin: Wydawnictwo Politechniki Lubelskiej.

Karczmarek, P., Kiersztyn, A., and Pedrycz, W. (2017a). "An evaluation of fuzzy measure for face recognition," in *International Conference on Artificial Intelligence and Soft Computing*. (Cham: Springer), 668–676. doi: 10.1109/TSMCB.2012.2185693

Karczmarek, P., Kiersztyn, A., and Pedrycz, W. (2017b). On developing Sugeno fuzzy measure densities in problems of face recognition. *Int. J. Mach. Intell. Sens. Signal Process.* 2, 80–96. doi: 10.1504/ijmissp.2017.088185

Karczmarek, P., Kiersztyn, A., and Pedrycz, W. (2018). Generalized choquet integral for face recognition. *Int. J. Fuzzy Syst.* 20, 1047–1055. doi: 10.1007/s40815-017-0355-5

Karczmarek, P., Kiersztyn, A., and Pedrycz, W. (2019b). "Generalizations of aggregation functions for face recognition," in *International Conference on Artificial Intelligence and Soft Computing*. (Cham: Springer), 182–192. doi: 10.1007/978-3-030-20915-5_17

Karczmarek, P., Pedrycz, W., Kiersztyn, A., and Dolecki, M. (2019a). A comprehensive experimental comparison of the aggregation techniques for face recognition. *Ira. J. Fuzzy Syst.* 16, 1–19.

Karczmarek, P., Pedrycz, W., Reformat, M., and Akhoundi, E. (2014). A study in facial regions saliency: a fuzzy measure approach. *Soft Comput.* 18, 379–391. doi: 10.1007/s00500-013-1064-0

Keefe, R. S. (2019). Why are there no approved treatments for cognitive impairment in schizophrenia? *World Psychiatry* 18, 167–168.

Klauser, P., Baker, S. T., Cropley, V. L., Bousman, C., Fornito, A., Cocchi, L., et al. (2017). White matter disruptions in schizophrenia are spatially widespread and topologically converge on brain network hubs. *Schizophr. Bull.* 43, 425–435. doi: 10.1093/schbul/sbw100

Klement, E. P., Mesiar, R., and Pap, E. (2000). *Triangular Norms*. Dordrecht: Springer.

Krukow, P., Jonak, K., Grochowski, C., Plechawska-Wójcik, M., and Karakuła-Juchnowicz, H. (2020). Resting-state hyperconnectivity within the default mode network impedes the ability to initiate cognitive performance in first-episode schizophrenia patients. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 102:109959. doi: 10.1016/j.pnpbp.2020.109959

Krukow, P., Jonak, K., Karakuła-Juchnowicz, H., Podkowiński, A., Jonak, K., Borys, M., et al. (2018). Disturbed functional connectivity within the left prefrontal cortex and sensorimotor areas predicts impaired cognitive speed in patients with first-episode schizophrenia. *Psychiatry Res.* 275, 28–35. doi: 10.1016/j.pscychresns.2018.03.001

Krukow, P., Karakuła-Juchnowicz, H., Juchnowicz, D., Morylowska-Topolska, J., Flis, M., and Jonak, K. (2017). Processing speed is associated with differences in IQ and cognitive profiles between patients with schizophrenia and their healthy siblings. *Nordic J. Psychiatry* 71, 33–41. doi: 10.1080/08039488.2016.1204469

Kwak, K. C., and Pedrycz, W. (2004). Face recognition using fuzzy integral and wavelet decomposition method. *IEEE Trans. Syst. Man Cybernet. Part B* 34, 1666–1675. doi: 10.1109/tsmcb.2004.827609

Kwak, K. C., and Pedrycz, W. (2005). Face recognition: a study in information fusion using fuzzy integral. *Pattern Recognit. Lett.* 26, 719–733.

Liu, H., Zhang, T., Ye, Y., Pan, C., Yang, G., Wang, J., et al. (2017). A data driven approach for resting-state EEG signal classification of schizophrenia with control participants using random matrix theory. *arXiv* [preprint], Avaliable online at: https://arxiv.org/abs/1712.05289 (accessed June 30, 2021).

Lucca, G., Sanz, J. A., Dimuro, G. P., Bedregal, B., Asiain, M. J., Elkano, M., et al. (2017). CC-integrals: Choquet-like copula-based aggregation functions and its application in fuzzy rule-based classification systems. *Knowl. Based Syst.* 119, 32–43. doi: 10.1016/j.knosys.2016.12.004

Lucca, G., Sanz, J. A., Dimuro, G. P., Bedregal, B., and Bustince, H. (2016). "Pre-aggregation functions constructed by CO-integrals applied in classification problems," in *Proceedings of IV CBSF*. (New York, NY: AMC), 1–11.

Lucca, G., Sanz, J. A., Dimuro, G. P., Bedregal, B., Mesiar, R., Kolesárová, A., et al. (2015). "The notion of pre-aggregation function," in *International Conference on Modeling Decisions for Artificial Intelligence*. (Cham: Springer), 33–41. doi: 10.1007/978-3-319-23240-9_3

Lucca, G., Vargas, R., Dimuro, G. P., Sanz, J., Bustince, H., and Bedregal, B. (2014). "Analysing some t-norm-based generalizations of the Choquet integral for different fuzzy measures with an application to fuzzy rule-based classification systems," in *ENIAC 2014, Encontro Nac. Intelig. Artificial e Computacional*. (São Carlos: SBC), 508–513.

Nolte, G., Ziehe, A., Krämer, N., Popescu, F., and Müller, K. R. (2010). "Comparison of Granger causality and phase slope index," in *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008,* Vol. 6, 267–276.

Pantelis, C., Yücel, M., Bora, E., Fornito, A., Testa, R., Brewer, W. J., et al. (2009). Neurobiological markers of illness onset in psychosis and schizophrenia: the search for a moving target. *Neuropsychol. Rev.* 19, 385–398. doi: 10.1007/s11065-009-9114-1

Parvinnia, E., Sabeti, M., Jahromi, M. Z., and Boostani, R. (2014). Classification of EEG Signals using adaptive weighted distance nearest neighbor algorithm. *J. King Saud Univers. Comput. Inform. Sci.* 26, 1–6. doi: 10.1016/j.jksuci.2013.01.001

Phang, C. R., Noman, F., Hussain, H., Ting, C. M., and Ombao, H. (2019). A multi-domain connectome convolutional neural network for identifying schizophrenia from EEG connectivity patterns. *IEEE J. Biomed. Health Informat.* 24, 1333–1343. doi: 10.1109/JBHI.2019.2941222

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Rosen, W. G., Mohs, R. C., Johns, C. A., Small, N. S., Kendler, K. S., Horvath, T. B., et al. (1984). Positive and negative symptoms in schizophrenia. *Psychiatry Res.* 13, 277–284.

Rutkowska, D., Kurach, D., and Rakus-Andersson, E. (2020). "Face recognition with explanation by fuzzy rules and linguistic description," in *International Conference on Artificial Intelligence and Soft Computing*. (Cham: Springer), 338–350.

Sabeti, M., Katebi, S. D., Boostani, R., and Price, G. W. (2011). A new approach for EEG signal classification of schizophrenic and control participants. *Expert Syst. Appl.* 38, 2063–2071. doi: 10.1016/j.eswa.2010.07.145

Sabeti, M., Sadreddini, M., and Price, G. (2007). "Fuzzy accuracy-based classifier systems for EEG classification of schizophrenic patients," in *In First Joint Congress on Fuzzy and Intelligent Systems Ferdowsi*. (Iran: University of Mashhad), 29–31.

Sheffield, J. M., Repovs, G., Harms, M. P., Carter, C. S., Gold, J. M., MacDonald, A. W. III, et al. (2015). Fronto-parietal and cingulo-opercular network integrity and cognition in health and schizophrenia. *Neuropsychologia* 73, 82–93. doi: 10.1016/j.neuropsychologia.2015.05.006

Shen, H., Wang, L., Liu, Y., and Hu, D. (2010). Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI. *Neuroimage* 49, 3110–3121. doi: 10.1016/j.neuroimage.2009.11.011

Shim, M., Hwang, H. J., Kim, D. W., Lee, S. H., and Im, C. H. (2016). Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features. *Schizophr. Res.* 176, 314–319. doi: 10.1016/j.schres.2016.05.007

Shim, M., Kim, D. W., Lee, S. H., and Im, C. H. (2014). Disruptions in small-world cortical functional connectivity network during an auditory oddball paradigm task in patients with schizophrenia. *Schizophr. Res.* 156, 197–203. doi: 10.1016/j.schres.2014.04.012

Silvana, M., Akbar, R., and Audina, M. (2018). "Development of classification features of mental disorder characteristics using the fuzzy logic Mamdani method," in *2018 International Conference on Information Technology Systems and Innovation (ICITSI)*. (Bandung: IEEE), 410–414.

Skudlarski, P., Jagannathan, K., Anderson, K., Stevens, M. C., Calhoun, V. D., Skudlarska, B. A., et al. (2010). Brain connectivity is not only lower but different in schizophrenia: a combined anatomical and functional approach. *Biol. Psychiatry* 68, 61–69. doi: 10.1016/j.biopsych.2010.03.035

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* 1:e42. doi: 10.1371/journal.pcbi.0010042

Stam, C. J., Tewarie, P., Van Dellen, E., van Straaten, E. C., Hillebrand, A., and Van Mieghem, P. (2014). The trees and the forest: characterization of complex brain networks with minimum spanning trees. *Int. J. Psychophysiol.* 92, 129–138. doi: 10.1016/j.ijpsycho.2014.04.001

Sugeno, M. (1974). *Theory of fuzzy Integral and Its Applications*. Dissertation, Tokyo: Tokyo Institute of Technology.

Szöke, A., Trandafir, A., Dupont, M. E., Meary, A., Schürhoff, F., and Leboyer, M. (2008). Longitudinal studies of cognition in schizophrenia: meta-analysis. *Br. J. Psychiatry* 192, 248–257. doi: 10.1192/bjp.bp.106.029009

Tewarie, P., van Dellen, E., Hillebrand, A., and Stam, C. J. (2015). The minimum spanning tree: an unbiased method for brain network analysis. *Neuroimage* 104, 177–188. doi: 10.1016/j.neuroimage.2014.10.015

Uhlhaas, P. J. (2013). Dysconnectivity, large-scale networks and neuronal dynamics in schizophrenia. *Curr. Opin. Neurobiol.* 23, 283–290. doi: 10.1016/j.conb.2012.11.004

Van Dellen, E., Bohlken, M. M., Draaisma, L., Tewarie, P. K., van Lutterveld, R., Mandl, R., et al. (2016). Structural brain network disturbances in the psychosis spectrum. *Schizophr. Bull.* 42, 782–789. doi: 10.1093/schbul/sbv178

Van Den Heuvel, M. P., and Fornito, A. (2014). Brain networks in schizophrenia. *Neuropsychol. Rev.* 24, 32–48.

van den Heuvel, M. P., and Sporns, O. (2013). Network hubs in the human brain. *Trends Cogn. Sci.* 17, 683–696. doi: 10.1016/j.tics.2013.09.012

Yager, R. R., and Kacprzyk, J. (eds) (2012). *The Ordered Weighted Averaging Operators: Theory and Applications*. Berlin: Springer Science & Business Media.

Yan, H., Tian, L., Wang, Q., Zhao, Q., Yue, W., Yan, J., et al. (2015). Compromised small-world efficiency of structural brain networks in schizophrenic patients and their unaffected parents. *Neurosci. Bull.* 31, 275–287. doi: 10.1007/s12264-014-1518-0

Zalesky, A., Fornito, A., and Bullmore, E. T. (2010). Network-based statistic: identifying differences in brain networks. *Neuroimage* 53, 1197–1207. doi: 10.1016/j.neuroimage.2010.06.041

Zalesky, A., Fornito, A., Seal, M. L., Cocchi, L., Westin, C. F., Bullmore, E. T., et al. (2011). Disrupted axonal fiber connectivity in schizophrenia. *Biol. Psychiatry* 69, 80–89.

Zhu, Q., Huang, J., and Xu, X. (2018). Non-negative discriminative brain functional connectivity for identifying schizophrenia on resting-state fMRI. *Biomed. Eng. Online* 17, 1–15. doi: 10.1186/s12938-018-0464-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 16 March 2022
doi: 10.3389/fninf.2021.802938

Counterfactual Explanation of Brain Activity Classifiers Using Image-To-Image Transfer by Generative Adversarial Network

Teppei Matsui[1,2*†], *Masato Taki*[3†], *Trung Quang Pham*[4], *Junichi Chikazoe*[4,5] and *Koji Jimura*[6]

[1] Department of Biology, Okayama University, Okayama, Japan, [2] JST-PRESTO, Japan Science and Technology Agency, Tokyo, Japan, [3] Graduate School of Artificial Intelligence and Science, Rikkyo University, Tokyo, Japan, [4] Supportive Center for Brain Research, National Institute for Physiological Sciences, Okazaki, Japan, [5] Araya Inc., Tokyo, Japan, [6] Department of Biosciences and Informatics, Keio University, Yokohama, Japan

Deep neural networks (DNNs) can accurately decode task-related information from brain activations. However, because of the non-linearity of DNNs, it is generally difficult to explain how and why they assign certain behavioral tasks to given brain activations, either correctly or incorrectly. One of the promising approaches for explaining such a black-box system is counterfactual explanation. In this framework, the behavior of a black-box system is explained by comparing real data and realistic synthetic data that are specifically generated such that the black-box system outputs an unreal outcome. The explanation of the system's decision can be explained by directly comparing the real and synthetic data. Recently, by taking advantage of advances in DNN-based image-to-image translation, several studies successfully applied counterfactual explanation to image domains. In principle, the same approach could be used in functional magnetic resonance imaging (fMRI) data. Because fMRI datasets often contain multiple classes (e.g., multiple behavioral tasks), the image-to-image transformation applicable to counterfactual explanation needs to learn mapping among multiple classes simultaneously. Recently, a new generative neural network (StarGAN) that enables image-to-image transformation among multiple classes has been developed. By adapting StarGAN with some modifications, here, we introduce a novel generative DNN (counterfactual activation generator, CAG) that can provide counterfactual explanations for DNN-based classifiers of brain activations. Importantly, CAG can simultaneously handle image transformation among all the seven classes in a publicly available fMRI dataset. Thus, CAG could provide a counterfactual explanation of DNN-based multiclass classifiers of brain activations. Furthermore, iterative applications of CAG were able to enhance and extract subtle spatial brain activity patterns that affected the classifier's decisions. Together, these results demonstrate that the counterfactual explanation based on image-to-image transformation would be a promising approach to understand and extend the current application of DNNs in fMRI analyses.

Keywords: fMRI, deep learning, explainable AI, decoding, generative neural network, counterfactual explanation

OPEN ACCESS

Edited by:
Itir Onal Ertugrul, Tilburg University, Netherlands

Reviewed by:
Shijie Zhao, Northwestern Polytechnical University, China
Rufin VanRullen, Centre National de la Recherche Scientifique (CNRS), France

***Correspondence:**
Teppei Matsui
tematsui@okayama-u.ac.jp

†These authors share first authorship

Received: *27 October 2021*
Accepted: *21 December 2021*
Published: *16 March 2022*

Citation:
Matsui T, Taki M, Pham TQ, Chikazoe J and Jimura K (2022) Counterfactual Explanation of Brain Activity Classifiers Using Image-To-Image Transfer by Generative Adversarial Network. Front. Neuroinform. 15:802938. doi: 10.3389/fninf.2021.802938

INTRODUCTION

Recent studies demonstrated promising results of the deep neural network (DNN) (LeCun et al., 2015) for decoding cognitive or behavioral information from brain activity images as observed with functional magnetic resonance imaging (fMRI) (Wang et al., 2020; Tsumura et al., 2021). However, despite these promising results, further applications of DNN to fMRI data could be limited due to its poor interpretability. Because of its highly non-linear and complex processing, it is often difficult to interpret what features of a given input led to the DNN's decision (Dong et al., 2019). For example, in the case of brain activity decoding, even though the DNN can accurately assign brain activations to a particular task, it is difficult to pinpoint which patterns of brain activations were important for the DNN's decisions. Such interpretability would be even more important when the DNN's decoding is incorrect. Gradient-based visualization methods, such as Grad-CAM (Selvaraju et al., 2020), are frequently used to highlight image regions potentially relevant for the DNN's decision [see Tsumura et al. (2021) for an application in neuroimaging]. However, several limitations of the gradient-based methods, such as high numbers of false positives (Eitel and Ritter, 2019), have been reported. Thus, alongside improving the gradient-based methods (Chattopadhay et al., 2018), it would be beneficial to explore alternative approaches for interpreting the inner workings of DNNs (Adadi and Berrada, 2018).

Counterfactual explanation is one of the major approaches for explaining DNN's inner working (Goyal et al., 2019; Wang and Vasconcelos, 2020). To explain how the decision on a given data was made, counterfactual explanation uses artificial data ("counterfactuals") that are generated from the real data but targeted to an unreal outcome (decision). By comparing the DNN's decision on the real data and the counterfactual, one can deduce explanations of the decisions made by the DNN. For example, we consider a case in which a brain activity classifier incorrectly assigns a gambling task to a brain activation produced in a motor task (**Figure 1A**). We consider a minimal transformation of the original brain activation to a counterfactual

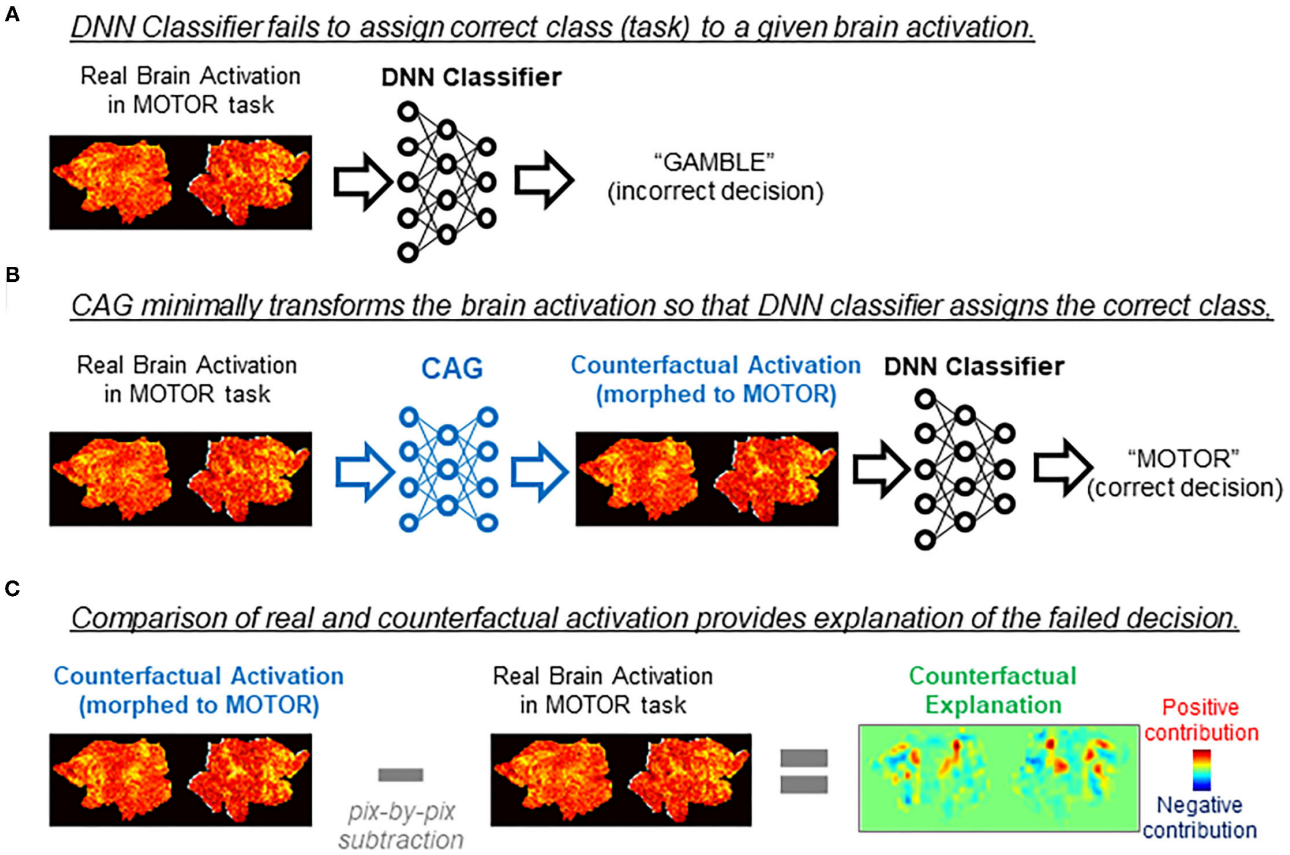


FIGURE 1 | Applications of counterfactual explanation in fMRI. The example illustrates an application of counterfactual explanation to a misclassification by a DNN classifier. **(A)** In this example, a DNN classifier incorrectly assigned EMOTION to a map of brain activation obtained in a MOTOR task. Because of the black-box nature of the DNN classifier, it is difficult to explain why the misclassification occurred. **(B)** A generative neural network for counterfactual brain activation (CAG) minimally transforms the real brain activation in **(A)** so that the DNN classifier now assigns MOTOR to the morphed activation (counterfactual activation). **(C)** Counterfactual explanation of misclassification in **(A)** can be obtained by taking the difference between the real activation and the counterfactual activation. In this example, the real brain activation would have been classified (correctly) as MOTOR if red (blue) brain regions in the counterfactual explanation had been more (less) active.

activation that is classified (correctly) as a motor task activation by the DNN classifier (**Figure 1B**). By directly comparing the original brain activation and the counterfactual activation, one can explain the classifier's decision by making a statement such as "This brain activation map would have been correctly classified to the gambling task if brain areas X and Y had been activated." (**Figure 1C**). As in this example, counterfactual explanation can provide intuitive explanations of a black-box decision system without opening the black-box, which is a critical aspect of the technique.

Although the generation of counterfactuals for high-dimensional data such as natural images and medical images had been difficult, recent advancement in DNN-based image generation has made counterfactual explanation applicable to these domains. For natural images, several studies have successfully used counterfactual explanation to explain the behavior of DNN-based image classifiers (Chang et al., 2019; Liu et al., 2019; Singla et al., 2020; Zhao, 2020). In medical image analyses, counterfactual explanation has also been applied to DNN-based classifiers of X-ray and structural MR images (Mertes et al., 2020; Pawlowski et al., 2020). However, to the best of our knowledge, counterfactual explanation has not been utilized for DNN-based classifiers of fMRI data. The lack of application to fMRI data may be due to the fact that commonly used fMRI dataset [such as data distributed by the Human Connectome Project (Van Essen et al., 2013)] usually contains data for multiple tasks. Because of this characteristic of fMRI dataset, unlike commonly used image generator that performs image-to-image transformation only between two classes, a generator of counterfactual brain activations needs to be able to transform the inputs to more than three classes.

Recently, the StarGAN (Choi et al., 2018) has enabled image-to-image transfer among multiple classes, thus opening the possibility to extend the application of counterfactual explanation to the multiclass fMRI dataset. In this study, based on the StarGAN model, we developed a generative neural network named counterfactual activation generator (CAG), which provides counterfactual explanations for a DNN-based classifier of brain activations observed with fMRI. This study aims to provide a proof of principle that counterfactual explanation can be applied to fMRI data and the DNN-based classifiers of brain activations. We demonstrated several applications of CAG. First, CAG could provide counterfactual explanations of correct classifications of the brain activations by DNN-based classifiers. Specifically, the counterfactual explanation highlighted the patterns of brain activations that were critical for the DNN classifier to assign the brain activations to particular tasks. Similarly, CAG could provide counterfactual explanations of incorrect classifications by DNN-based classifiers. Moreover, iterative application of CAG accentuated and extracted subtle image patterns in brain activations that could strongly affect the classifier's decisions. These results suggest that the image transfer-based methods, such as CAG, would be a powerful approach for interpreting and extending DNN-based fMRI analyses.

MATERIALS AND METHODS

Datasets

Training data were single-subject second-level z-maps obtained during the performance of seven behavioral tasks from the S1200 release of the Human Connectome Project ($N = 992$; HCP; http://www.humanconnectomeproject.org/) (Barch et al., 2013; Van Essen et al., 2013; Glasser et al., 2016). From each participant, statistical z-maps were obtained for activation contrasts for the emotional processing task (face vs. shape), the gambling task (reward vs. loss), the language processing task (story vs. math), the motor task (average of all motions), the relational processing task (relational processing vs. matching), the social cognition task (mental vs. random), and the N-back working memory task (2-back vs. 0-back). For brevity, the seven tasks are denoted as follows: (1) EMOTION, (2) GAMBLING, (3) LANGUAGE, (4) MOTOR, (5) RELATIONAL, (6) SOCIAL, and (7) WORKING MEMORY (WM). We used gray-scaled flat 2D cortical maps (Glasser et al., 2016) provided from HCP for dimensional compatibility of images between VGG16-ImageNet and activation maps. The flattened maps were created using the Connectome Workbench (https://www.humanconnectome.org/software/connectome-workbench/) following a procedure described in (Tsumura et al., 2021) (**Figure 2A**).

DNN Classifier of Brain Activations

The DNN classifier of brain activations used in this study was adapted from our previous study (Tsumura et al., 2021) (**Figure 2B**). Briefly, the DNN classifier was based on VGG16 (Simonyan and Zisserman, 2015), with five convolution layers for extracting image features and two fully connected layers for classification of the seven tasks. Initial parameters of convolution layers were set to parameters pretrained with concrete object images provided from ImageNet (Simonyan and Zisserman, 2015) (http://www.image-net.org/). The VGG16/ImageNet model is capable of classifying concrete object images into 1,000 item categories. Importantly, it has been demonstrated that the pretrained model can learn novel image sets more efficiently than the non-trained model by tuning convolution and fully connected layers and fully connected layers only (Pan and Yang, 2010). Thus, the current analysis retrained the pretrained VGG16-ImageNet model to classify brain activation maps. To enable processing by generative neural networks described below, activation maps were spatially downsampled from 570 by 1,320 pixels to 50 by 140 pixels. Data were split into training data ($N = 4,730$) and validation data ($N = 518$) (note that some participants in the dataset did not complete all seven tasks). Training was conducted using the training data with ten-fold crossvalidation. Hyperparameters for the training were as follows: batch size, 10; epoch, 50; learning rate, 0.0001; optimizer, stochastic gradient descent (SGD); loss function, categorical crossentropy. Pixels outside of the brain were set to zero. Model training and testing were implemented using Keras (https://keras.io/) under a Tensorflow backend (https://www.tensorflow.org/). Five instances of the DNN classifier were trained for replication. All

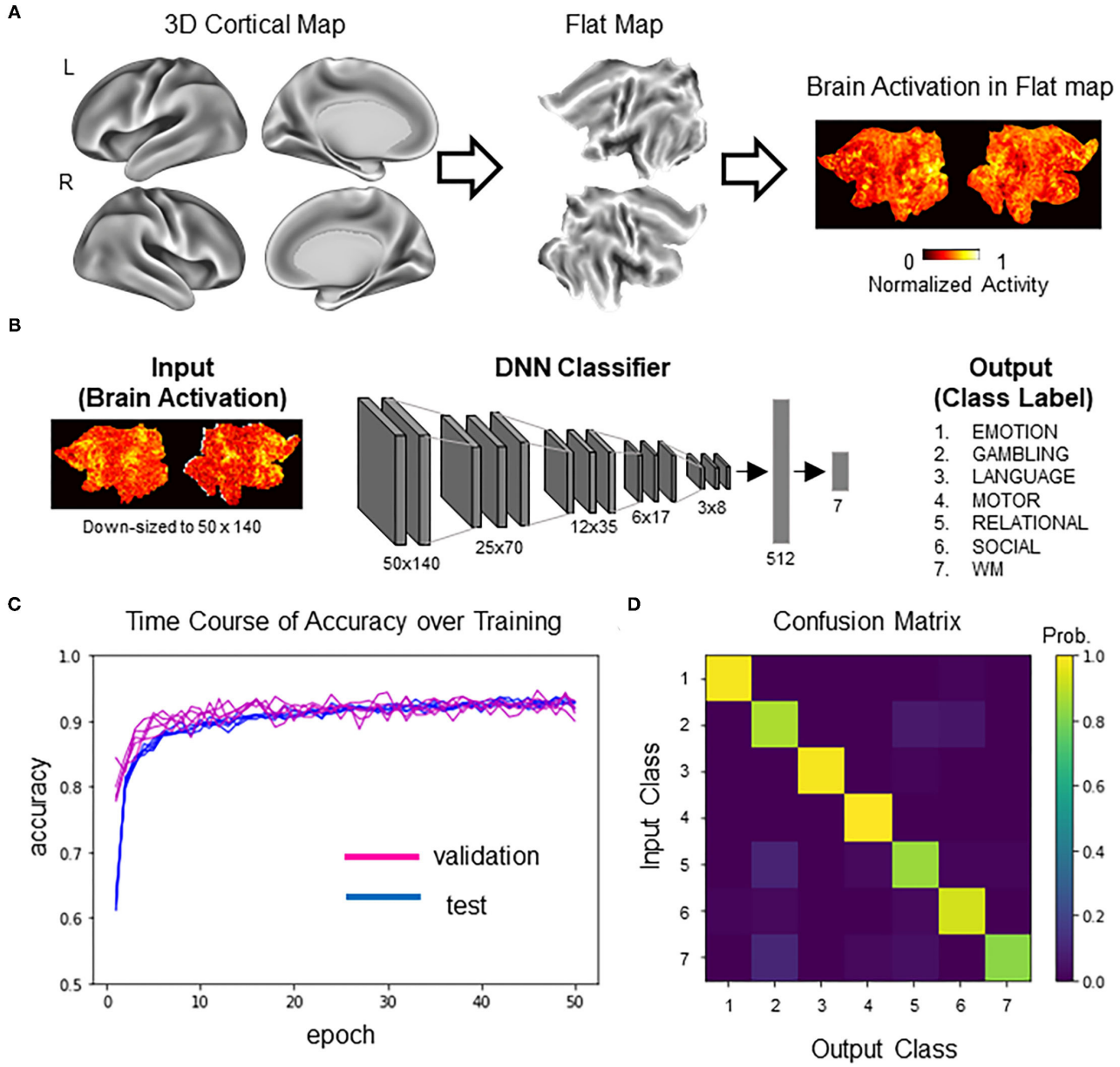


FIGURE 2 | DNN classifier for brain activity decoding. **(A)** Following the standard procedure developed by HCP (Glasser et al., 2016), neocortex in the two hemispheres was mapped to two cortical sheets. Each neocortical activity image was mapped to the two sheets, which was then input to the DNN classifier (for details, see Tsumura et al., 2021). **(B)** Model architecture of the DNN classifier. The input was a picture containing two sheets of cortical activations. The picture was downsampled for later processing by the generative neural network. The DNN classifier was a deep convolutional network similar to the one described in our previous study (Tsumura et al., 2021). The output of the DNN classifier was one-hot vectors representing seven behavioral tasks in the HCP dataset. **(C)** Training history of the transfer learning. Test accuracy (blue) and validation accuracy (magenta) are shown for five replicates. Note that the chance level is 14.3% (1/7). **(D)** Profile of the classifier's decision (confusion matrix) in the validation set.

parameters, including the training data, were the same for all the replicates.

Counterfactual Activation Generator (CAG)

We adopted the architecture of StarGAN (Choi et al., 2018), consisting of discriminator and generator, with a modification to add a new loss term for the DNN classifier (**Figure 3A**; see also **Supplementary Figure 1** for an illustration of our overall approach). Except for this addition of the new loss term (CAG loss), other parameters were the same as in the original StarGAN model. Briefly, the goal was to train a single generator that learns mapping among multiple classes (in this case, the seven HCP tasks). We regarded this generator as CAG. To achieve this, we trained CAG to transform a brain activation

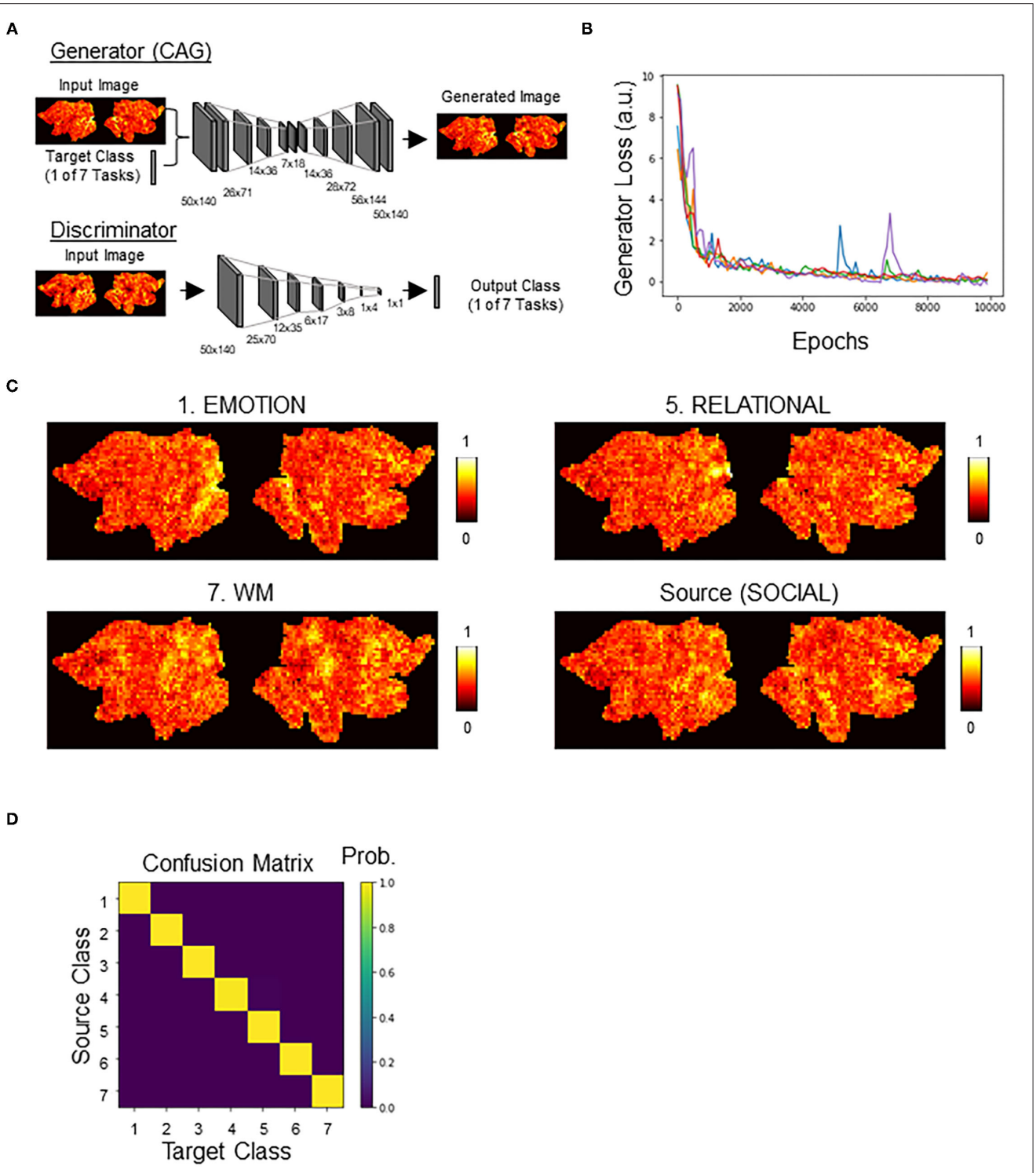


FIGURE 3 | Counterfactual activation generator (CAG). **(A)** Generator and discriminator architectures, see Methods for details. Networks were modified from StarGAN. Generator, once trained, served as CAG. Generator takes a combination of an image of brain activation and a one-hot label indicating the target class as an input. Generator outputs a counterfactual brain activation that is a minimal transform of the input brain activation toward the target class. Discriminator takes an activation map output by generator and outputs a one-hot label. Discriminator was cotrained with generator, as in StarGAN. **(B)** Time courses of the generator loss. Different colors indicate different replicates ($n = 5$). **(C)** Representative counterfactual activations generated by CAG. All counterfactual activations were generated from the source activation. See **Supplementary Figure 3** for transformation to all categories. **(D)** Confusion matrix showing the classifier's decision profile on counterfactual activations.

$\boldsymbol{x}$ with class-label y (source class) to a perturbation toward y^c (target class), such that $CAG\left(x\,,\,y^c\right) \rightarrow x^c$. An auxiliary discriminator was introduced to allow a single discriminator to control multiple classes. Thus, the discriminator produced the probability distributions over both the source and the target classes, $D : x \rightarrow \{D_{src}\,(x)\,,\,D_{cls}\,(x)\}$.

The loss terms were described as follows. Wasserstein loss ($\mathcal{L}_{wass}$) and gradient penalty loss ($\mathcal{L}_{gp}$) were included to make the generated brain activations indistinguishable from the real brain activations. $\mathcal{L}_{wass}$ between the real and counterfactual activations and $\mathcal{L}_{gp}$ were defined as follows:

$$\mathcal{L}_{wass} = \mathbb{E}_x\left[D_{src}\,(x)\right] - \mathbb{E}_{x,c}\left[D_{src}\,(CAG\,(x,c))\right]$$
$$\mathcal{L}_{gp} = \mathbb{E}_{\hat{x}}\left[\left(\parallel \nabla_{\hat{x}}\, D_{src}\left(\hat{x}\right) \parallel_2 - 1\right)^2\right]$$

where $\hat{x}$ was sampled uniformly along a straight line between a pair of a real and a generated activation.

Domain classification loss was included to ensure that the transformed activation was properly classified as the target class. We considered two types of objectives. The first one is a domain classification loss of real activations used to optimize discriminator ($\mathcal{L}^r_{cls} = \mathbb{E}_{x,c'}[-\log\, D_{src}\left(c'|x\right)]$). The second one is a domain classification loss of fake activations used to optimize CAG ($\mathcal{L}^f_{cls} = \mathbb{E}_{x,c}[-\log\, D_{src}\left(c|CAG(x,c)\right)]$). This loss term forced CAG to generate activations that could be classified as the target classes.

Reconstruction loss ($\mathcal{L}_{rec}$) was defined using the cycle consistency loss (Kim et al., 2017; Zhu et al., 2017) as follows,

$$\mathcal{L}_{rec} = \mathbb{E}_{x,c,c'}\left[\left(\parallel x - \mathrm{CAG}(\mathrm{CAG}\,(x,\mathrm{c})\,,c')\parallel_1\right)\right]$$

where CAG tried to reconstruct the original activation from the transformed activation.

Additionally, we included a loss term for the DNN classifier ($\mathcal{L}_{cnn}$) to force the mappings learned by CAG to be aligned with the classifier's decisions. This loss term was calculated using categorical crossentropy over the fake activations.

The total losses for discriminator ($\mathcal{L}_D$) and the CAG loss ($\mathcal{L}_G$) were defined using the loss terms as follows:

$$\mathcal{L}_D = \mathcal{L}^r_{wass} + \mathcal{L}^f_{wass} + \lambda_{gp}\mathcal{L}_{gp} + \lambda_{cls}\mathcal{L}^r_{cls}$$
$$\mathcal{L}_G = \mathcal{L}_{wass} + \lambda_{cls}\mathcal{L}^f_{cls} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cnn}\mathcal{L}_{cnn}$$

where $\mathcal{L}^r_{wass}$ and $\mathcal{L}^f_{wass}$ stand for Wasserstein loss for real and fake activations, respectively. We used the same hyperparameters and procedures used in the original StarGAN model, except for λ_{cnn} which was newly introduced in CAG. Instance normalization was used for the generator, but no normalization was used for the discriminator. The generator network consisted of three convolutional layers for downsampling, followed by two convolutional layers (replacing two residual blocks in the original StarGAN model), which was intern followed by four convolutional layers for up sampling. We used $\lambda_{gp} = 10$, $\lambda_{cls} = 1$, $\lambda_{rec} = 10$, and $\lambda_{cnn} = 1$ for all experiments. All models were trained using Adam (Kingma and Ba, 2014), with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Training was done using the training data with ten-fold crossvalidation for 10,000 epochs. Batch size and learning rate were set to 16 and 0.0001, respectively, in all experiments. The code for CAG is available upon reasonable request to the corresponding author.

Counterfactual Explanation of Correctly and Incorrectly Classified Images

Counterfactual explanation of correctly classified images was performed on the correctly classified brain activations ($N = 478$ out of 518 that were not used in the classifier training). Each counterfactual explanation was set to explain "Why this activation was correctly classified as class (task) A instead of class B?" To do this, the original brain activation was transformed by CAG toward class B. Counterfactual explanation was obtained by pixel-by-pixel subtraction of the original activation from the counterfactual activation. As for counterfactual explanation of correct classifications, counterfactual activations were obtained by transforming the correctly classified activations to one of the randomly chosen incorrect classes.

To quantitatively evaluate the effectiveness of counterfactual explanations, we conducted two analyses. In the first analysis, we perturbed image transformation by CAG at various levels and examined its effect on the classifier's decisions. For the perturbation, pixels in each counterfactual explanation whose values were below a chosen percentile threshold (α) were set to zero (CE_α). Then, the perturbed counterfactual explanation was added back to the original activation ($Activation_{original}$) as follows:

$$Activation_{new} = Activation_{original} + CE_\alpha$$

The resulting activation ($Activation_{new}$) was normalized to have minimum and maximum values of zero and one, respectively, and then input to the DNN classifier. The percentile threshold (α) took values ranging from 0 to 100% with a 20% step. Note that $Activation_{new}$ is equal to the counterfactual activation and $Activation_{original}$ when α equals to 0 and 100%, respectively. In the second analysis, each counterfactual explanation was compared with a "control explanation," which was calculated as the difference between the true class's average activations and the target class used for the transformation (ΔAve). The control explanation was added to the original activation ($Activation_{original}$) as follows,

$$Activation_{new} = Activation_{original} + \Delta Ave \times \kappa$$

The resulting activation ($Activation_{new}$) was normalized to have minimum and maximum values of zero and one, respectively, and then input to the DNN classifier. The parameter for mixing (κ) took values ranging from 0 to 5 at with a 0.1 step and was adjusted individually for each control explanation to maximize the total number of cases classified to the target class used for transformation.

Counterfactual explanation of incorrectly classified images was performed similarly on each incorrectly classified brain activation ($N = 40$). Each counterfactual explanation was set

to explain "Why this activation was incorrectly classified as class (task) B instead of class A?" To do this, the original brain activation was transformed by CAG toward the true class A. Counterfactual explanation was obtained by pixel-by-pixel subtraction of the original activation from the counterfactual activation. The two quantitative analyses for the counterfactual explanation of correct classifications were similarly applied to the counterfactual explanation of incorrect classifications. In these analyses, the target class for the image transformation by CAG was set to the correct classes (instead of randomly chosen classes in the case of correct classification).

Counterfactual Exaggeration and Feature Extraction

Counterfactual exaggeration (Singla et al., 2020) was performed by iteratively transforming a real brain activation toward one class. We performed up to eight iterations. Feature extraction was done by subtracting the third iteration from the eighth iteration. To quantitatively evaluate the extracted feature, the feature was added to each activation in the validation set ($N = 518$), and then, the summed image was input to the DNN classifier. For 12 extracted features from randomly chosen activations, the percent of activations assigned to the added feature's class were calculated.

RESULTS

DNN Classifier Decoded Task Information From Brain Activity With High Accuracy

We first trained a DNN classifier that was used as the target for counterfactual explanation. Brain activations were converted to flattened maps, which were then input to the DNN classifier (**Figure 2A**). The DNN classifier was based on VGG16 pretrained on the ImageNet dataset (Tsumura et al., 2021) (**Figure 2B**). The pretrained DNN classifier was trained to classify brain activation maps using transfer learning (Pan and Yang, 2010). After 50 epochs of training, the DNN classifier reached ∼92% of classification accuracy for the held-out validation data. Similar results were obtained for a total of five replicates, suggesting high reproducibility (**Figure 2C**). **Figure 2D** shows the confusion matrix showing the classifier's decision profile (see also **Supplementary Table 1** for exact values). Similar confusion matrices were obtained for all the replicates (data not shown). These results suggest that DNN classifiers could accurately decode task information from individual brain activations.

CAG Generated Counterfactual Activations Were Realistic and Fooled the Classifiers

We next trained a generative neural network (CAG) for counterfactual explanations of the DNN classifier's decisions. For this, we adopted, with modifications, the architecture of StarGAN (Choi et al., 2018) that can perform image-to-image transformation among multiple classes. Two DNNs, generator (CAG) and discriminator, were simultaneously trained (**Figure 3A**; **Supplementary Figure 1**). By including the classification loss by the DNN classifier, CAG was trained to simultaneously fool both the discriminator and the DNN classifier (**Supplementary Figure 1**; see Methods for details). Throughout the training, the generator loss, which is a good indicator of the quality of the generated image (Arjovsky et al., 2017), consistently decreased toward zero and plateaued around 10,000 epochs of training (data for five replicates are shown in **Figure 3B**; see also **Supplementary Figure 2** for time courses of all the loss terms). After the training, CAG could transform a real brain activation into a counterfactual brain activation that was visually indistinguishable from the real activations (**Figure 3C**; **Supplementary Figure 3**). The training of the CAG was also designed such that the CAG transformed an input activation map to any of the seven classes and fool the DNN classifier. Thus, the trained DNN assigned the targeted class to the counterfactual activations at almost 100% accuracy (**Figure 3D**; **Table 1**). These results suggest that CAG fulfilled the goal of generating counterfactual brain activations that were not only visually realistic but also fooled the DNN classifier.

Counterfactual Explanation of Misclassification by DNN Classifiers

Using CAG, we first conducted counterfactual explanation of the classifier's correct decisions. Specifically, we tried to visualize the pattern of brain activation that led the classifier to assign the correct class but not another (incorrect) class (**Figure 4A**). In the first example, brain activations correctly classified as MOTOR by the DNN classifier were examined (**Figure 4B**). We asked why these activations were not classified as EMOTION. To see this, a counterfactual activation was created by transforming each original activation toward EMOTION using CAG (**Figure 4C**). Then, the counterfactual explanation was obtained by taking the difference between the original and the counterfactual activations (**Figure 4D**). The positive and negative regions in the counterfactual explanation were the regions that had positive and negative influence, respectively, on the classifier's decision of assigning EMOTION but not MOTION to the counterfactual activation. In other words, the DNN classifier would have

TABLE 1 | Decision profile of DNN classifier on counterfactual activations.

CLASS	EMOTION	GAMBLING	LANGUAGE	MOTOR	RELATIONAL	SOCIAL	WM
$N_{correct}$ (%)	518 (100%)	518 (100%)	518 (100%)	512 (98.8%)	518 (100%)	518 (100%)	518 (100%)

Each image in the validation set (N = 518) was morphed toward one of the seven classes and then input to the DNN classifier.

classified the original activation as EMOTION if the positive regions in the counterfactual explanation had been more active (and the opposite for the negative regions).

To quantitatively evaluate the counterfactual explanation, we compared it against a difference between the population-averaged activations of EMOTION and MOTOR (average of 74 and 75 activations, respectively) (**Figure 4E**). Whereas the difference of average maps highlighted a small portion of brain areas in the occipital cortex, counterfactual explanation additionally found lateral temporal areas to be relevant. Because lateral temporal areas are known to be activated by emotional and facial processing (White et al., 2014; Glasser et al., 2016), it is reasonable to find these areas highlighted in the counterfactual explanation. The reason that only the occipital cortex was highlighted in the difference in average maps is most likely due to very high activation in this area in the average map of EMOTION compared to the average map of MOTOR. These differences between the counterfactual explanation and the explanation by the difference of averages can be understood as the difference between univariate and multivariate analyses (Jimura and Poldrack, 2012). The average map that is derived from the univariate analysis (pixel-based GLM) is affected by the choice of particular control conditions, and thus, the explanation by the difference of the averages would be affected by difference in the control conditions. In contrast, the counterfactual explanation can robustly detect relevant activations using spatial patterns of multiple pixels. Consistent with this idea, the counterfactual explanation, but not the difference of average maps, successfully highlighted the orbitofrontal areas implicated for emotional processing (**Figure 4D**) (Goodkind et al., 2012; Rolls et al., 2020).

The second example shows counterfactual explanation of why WM activations were not classified as LANGUAGE (**Supplementary Figure 4**). In this case, unlike the previous example, the difference of averages of WM and LANGUAGE highlighted large portions of the brain (red regions in **Supplementary Figure 4**; average of 75 and 73 maps, respectively, for WM and LANGUAGE). With such large and distributed areas being highlighted, it is difficult to pinpoint particular areas without arbitrary thresholding. In contrast, the counterfactual explanation highlighted distributed but much more localized brain areas (**Supplementary Figure 4C**). It is evident from the counterfactual explanation that activations in frontal and temporal brain areas would have been necessary to shift the DNN classifier's decision from WM to LANGUAGE.

Out of 478 correctly classified validation data, 476 counterfactual activations were classified as the targeted class. To test the robustness of the result against image corruption, we isolated the image components added by CAG (i.e., the difference between the counterfactual activation and the raw activation). Then, we perturbed the image components at different levels of percentile thresholds (α in **Table 2**. see methods) that were in turn added back to the raw activation. The effect of thresholding did not change the classification results when the bottom 20% of the image components were perturbed. The classification results were still above 25%, even when the bottom 60% of the image components were perturbed. The classification results were markedly degraded when the bottom 80% of the image components were perturbed. Thus, these results suggest that image modifications imposed by CAG were robust to perturbation in a large margin. To further assess the effectiveness of counterfactual explanation, we compared the classifier's response to counterfactual activations and control maps obtained by adding the original activation and the difference of average activations ("Control" in **Table 2**). Only four of the control maps were classified as the targeted class. Together, these results demonstrated that counterfactual explanations provided interpretable activation patterns that could not only explain the classifier's decisions but also robustly manipulate the classifier's decisions.

Note that the aim of the discussion here is not to infer cognitive tasks associated with the brain activation, a type of discussion considered as reverse inference (Poldrack, 2006). In this case, the cognitive tasks (i.e., classes) associated with the brain activations were entirely determined by the DNN classifier. The purpose of the discussion here is to interpret the counterfactual explanation in relation to existing knowledges about the brain activity. In the future, this type of discussion may be automated using applications such as Neurosynth (Yarkoni et al., 2011).

Counterfactual Explanation of Misclassification by DNN Classifiers

An important feature of counterfactual explanation is its ability to provide explanations to single cases of misclassification. We next demonstrated this in misclassifications by the DNN classifier (**Figure 5A**). For each case of misclassifications, the misclassified activation map was transformed toward the correct class by CAG. Then, the difference between the counterfactual activation and the real (misclassified) activation was calculated for the counterfactual explanation. In the first example, a brain activation in the EMOTION task was incorrectly classified as SOCIAL (**Figure 5B**). A counterfactual activation was obtained by transforming the real activation toward the correct class (EMOTION) (**Figure 5C**). Interestingly, the counterfactual explanation suggested that activations in the occipital regions were critically lacking for the DNN classifier to classify the original activation as EMOTION (**Figure 5D**). Because the occipital area is considered to process low-level visual information (Yamins et al., 2014), this occipital activation likely indicates bias in the dataset that used visual stimulus in the EMOTION task (Barch et al., 2013) rather than a brain activation related to emotional processing. Thus, counterfactual explanation revealed that this misclassification was likely due to the bias in the dataset, which was unintentionally learned by the DNN classifier.

As for a control analysis that can be compared with the counterfactual explanation, we calculated the difference between the misclassified (real) activation and the average activation of EMOTION (**Figure 5E**). Despite the similar global trend with the counterfactual explanation, the difference with the average showed a noisy pattern whose local peaks were difficult to find. Importantly, a peak in the occipital area was difficult to discern in the difference with the average. In

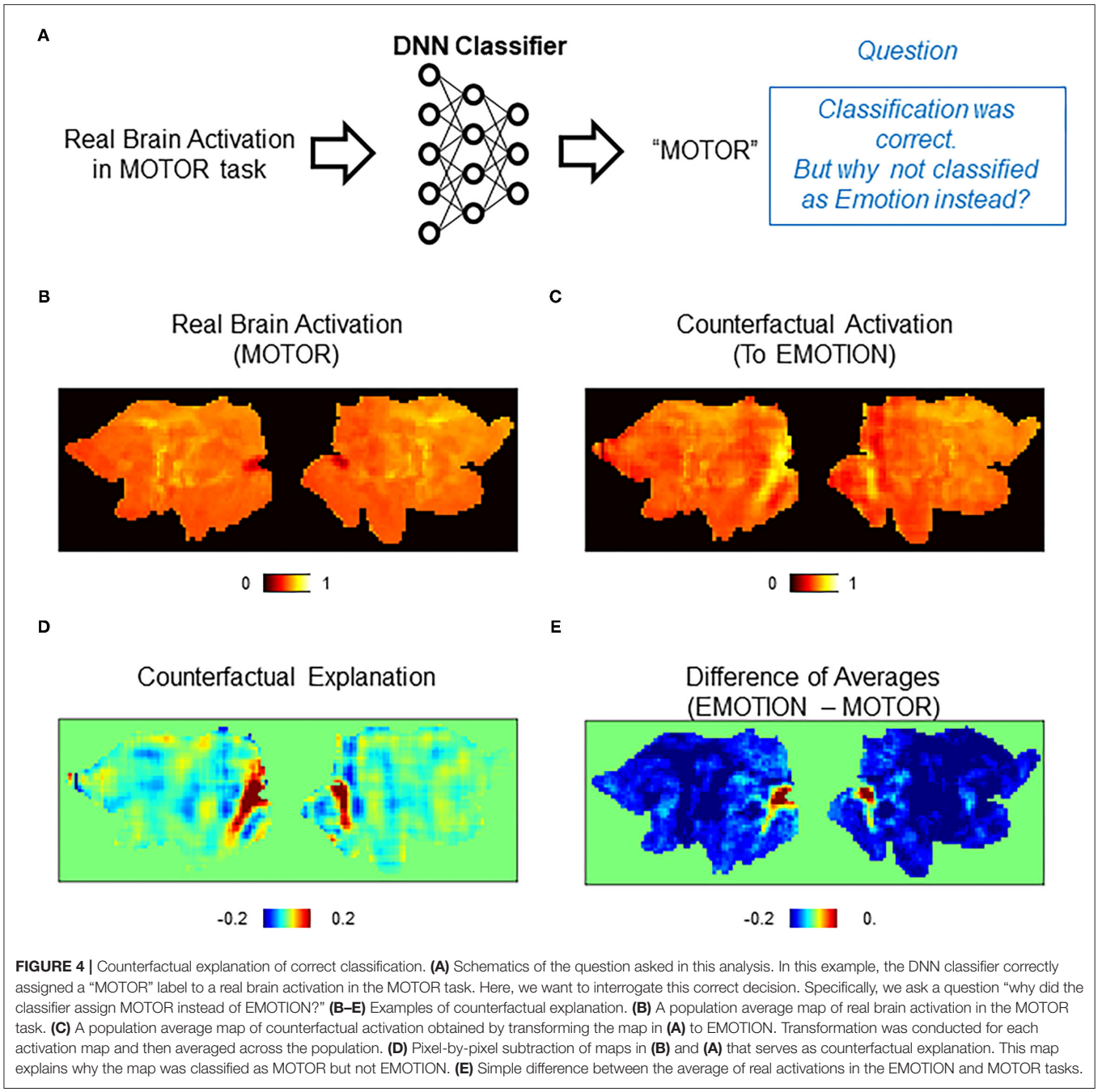


FIGURE 4 | Counterfactual explanation of correct classification. **(A)** Schematics of the question asked in this analysis. In this example, the DNN classifier correctly assigned a "MOTOR" label to a real brain activation in the MOTOR task. Here, we want to interrogate this correct decision. Specifically, we ask a question "why did the classifier assign MOTOR instead of EMOTION?" **(B–E)** Examples of counterfactual explanation. **(B)** A population average map of real brain activation in the MOTOR task. **(C)** A population average map of counterfactual activation obtained by transforming the map in **(A)** to EMOTION. Transformation was conducted for each activation map and then averaged across the population. **(D)** Pixel-by-pixel subtraction of maps in **(B)** and **(A)** that serves as counterfactual explanation. This map explains why the map was classified as MOTOR but not EMOTION. **(E)** Simple difference between the average of real activations in the EMOTION and MOTOR tasks.

the second example, we examined an activation in WM that was misclassified as GAMBLING (**Supplementary Figure 5A**). A counterfactual activation was obtained by transforming the real activation toward WM (**Supplementary Figure 5B**). As in the first example, the counterfactual explanation showed a pattern of brain activation with multiple identifiable peaks (**Supplementary Figure 5C**). In contrast, the difference with the average provided a noisier pattern whose local peaks were difficult to identify (**Supplementary Figure 5D**). These results demonstrated that counterfactual explanation can provide interpretable patterns of brain activations related to individual cases of misclassifications by the DNN classifier.

Next, we quantitatively assessed the counterfactual explanation of misclassifications. The DNN classifier assigned the correct classes to all the counterfactual activations that are equivalent to additions of the real (misclassified) activations and the counterfactual explanations (40 of 40 misclassified activations in the validation set). To assess the robustness of the results to

image perturbation, we conducted the same analysis that we used for the correct classification. In the case of misclassification, the DNN classifier assigned the correct classes in 80% of cases, even when the bottom 80% of the image components modified by CAG were perturbed (**Table 3**). This result suggests that only a small modification to the misclassified activation was necessary to shift the classifier's decision to the correct class.

As for the control analysis, for each misclassified activation, we calculated the control activation that is the sum of the misclassified activation and the difference of averages of the true class and the incorrectly assigned class. In contrast to counterfactual activations, only two of the control activations were classified as the true classes (**Table 3**). These results suggest that counterfactual explanation, but not the addition of the difference of average activations, captured the image transformation needed to correct the decisions of the DNN classifier.

Counterfactual Exaggeration Revealed Subtle Image Features Important for the Classifications by DNN

In addition to counterfactual explanations of correct and incorrect classifications, the deep image generator can perform "counterfactual exaggeration" to enhance and detect subtle image features exploited by DNNs (Singla et al., 2020). In counterfactual exaggeration, an image is iteratively transformed by the generator toward one class. This iterative image transformation enhances subtle image features exploited by DNNs. In a previous work, exaggerated images were used to discover a novel symptom of diabetic macular edema (Narayanaswamy et al., 2020). Inspired by these previous works, we next used CAG in counterfactual exaggeration to detect subtle features of brain activations exploited by the DNN classifier (**Figure 6A**). Interestingly, in some cases, iterative application of CAG revealed a texture-like feature in the image (**Figures 6B–D**). Such a texture-like feature was difficult to discern in the original activation (**Figure 6B**) but became evident as the counterfactual exaggeration was repeatedly applied (**Figures 6C,D**). The texture-like feature could be extracted by taking the difference in counterfactual activations with different numbers of iterations (**Figure 6E**).

Although the texture-like pattern did not appear in the same way as a real brain activation, it could nevertheless influence the classifier decisions. In fact, it has been suggested that DNNs are biased toward using textures for image classification (Geirhos et al., 2018). To quantitatively examine this point, we added the extracted features to randomly chosen real activations and then examined the resulting activations by the DNN classifier. **Figure 6F** and **Supplementary Figure 6** show examples of the extracted features and the real activations before and after the addition of the features. Note that differences between the appearance of activations before and after the addition of the features were subtle because the amplitudes of the extracted features were relatively small. Nevertheless, the addition of the extracted features caused the DNN classifier to (mis-)assign the activations the classes to which the exaggerations were targeted (**Figure 6G**). Misclassification to the targeted class occurred in 55.0 ± 26.1% of cases (mean ± standard deviation; N=12 extracted features; $p < 0.001$, sign rank test; see Methods for details). These results suggest that counterfactual exaggeration assisted by CAG was able to enhance and discover subtle image features that are exploited by the DNN classifier. The texture-like features likely represent image features relevant to adversarial vulnerability of the DNN classifier (Geirhos et al., 2018). Being able to detect and protect against such attacks is critical for future reliable applications of DNN-based brain decoders.

TABLE 2 | Decision of DNN classifier on counterfactual activations obtained from correctly classified brain activations.

Correct cases	**Counterfactual activations**						**Control**
	α = 0	**α = 20**	**α = 40**	**α = 60**	**α = 80**	**α = 100**	
$N_{correct}$ (%)	476 (99.6%)	470 (98.0%)	327 (68.4%)	125 (26.2%)	9 (1.9%)	0 (0.0%)	4 (0.8%)

Counterfactual activations were obtained from images correctly classified by the DNN classifier (N = 478). Each image was transformed to one of a number of randomly chosen incorrect classes. All counterfactual activations were classified as the targeted class by the DNN classifier (middle column). As for the control analysis, the difference of the average maps for targeted vs. original classes was added to each image. None of the control images were classified as the targeted class (right column).

TABLE 3 | Decision of the DNN classifier on counterfactual activations obtained from misclassified brain activations.

Incorrect cases	**Counterfactual activations**						**Control**
	α = 0	**α = 20**	**α = 40**	**α = 60**	**α = 80**	**α = 100**	
$N_{correct}$ (%)	40 (100%)	40 (100%)	39 (97.5%)	38 (95.0%)	24 (60.0%)	0 (0.0%)	2 (0.5%)

Counterfactual activations were obtained from images originally misclassified by the DNN classifier (N = 40). All of the counterfactual activations were correctly classified by the DNN classifier after transformation by CAG (middle column). As for the control analysis, the difference of the average maps for incorrect and correct classes was added to each misclassified image. None of the control images were classified as the targeted class (right column).

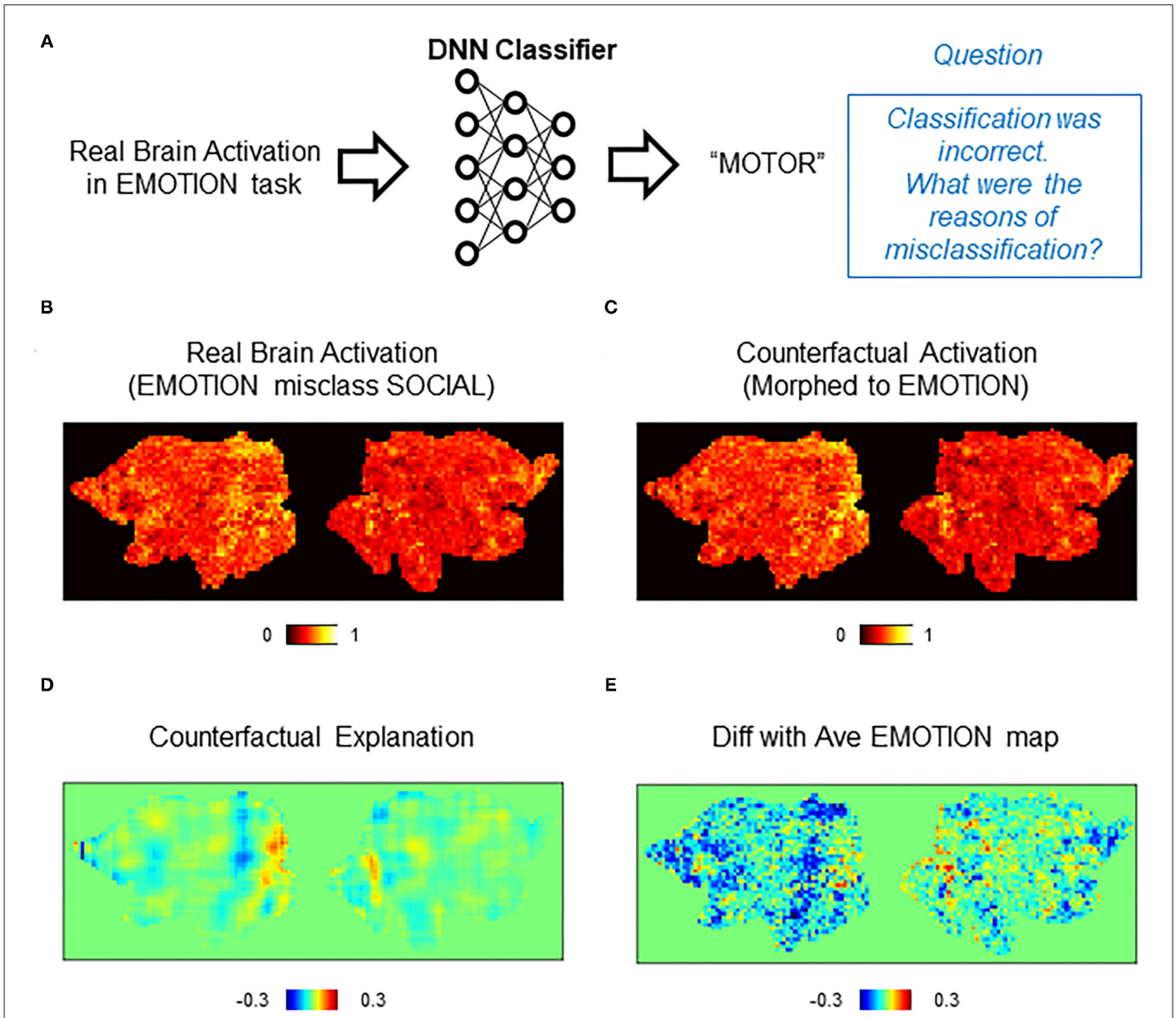


FIGURE 5 | Counterfactual explanation of incorrect classification. **(A)** Schematics of the question asked in this analysis. In this example, the DNN classifier incorrectly assigned a "SOCIAL" label to a real brain activation in the EMOTION task. Here, we want to interrogate this incorrect decision. Specifically, we ask a question "why did the classifier (incorrectly) assign EMOTION instead of SOCIAL?" **(B–E)** Example of counterfactual explanation. **(B)** A single brain activation map for EMOTION that was incorrectly classified as SOCIAL by the DNN classifier. **(C)** A map of counterfactual activation obtained by transforming the map in **(A)** to SOCIAL. **(D)** Pixel-by-pixel subtraction of maps in **(B)** and **(A)** that serves as counterfactual explanation. This map explains why the map was incorrectly classified as SOCIAL but not EMOTION. **(E)** Simple difference between the average of real activations in the EMOTION and the single activation map for SOCIAL shown in **(A)**.

DISCUSSION

In this study, we provided a proof-of-principle of application of counterfactual explanations from a generative model to understand the decisions of a DNN trained to decode task information from brain activation data. In the field of computer vision, such explanation is often conducted with saliency maps that highlight regions in the input that were important for the decisions of the DNN (Selvaraju et al., 2020). A recent neuroimaging study also used saliency maps to interpret decisions made by DNNs (Tsumura et al., 2021). A limitation of this approach is that the regions highlighted in saliency maps are not necessarily causally related to the decisions of the DNN (Eitel and Ritter, 2019). Hence, user of saliency maps needs to perform an additional interpretation of why the highlighted areas are important for the DNN's decisions (Mertes et al., 2020). In contrast to saliency maps, counterfactual explanation explains why the actual decision was made instead of another one. By creating a slightly modified version of the input that leads another decision by the DNN, counterfactual explanation provides a different kind of explanatory information which helps to interpret saliency maps. Thus, future neuroimaging studies

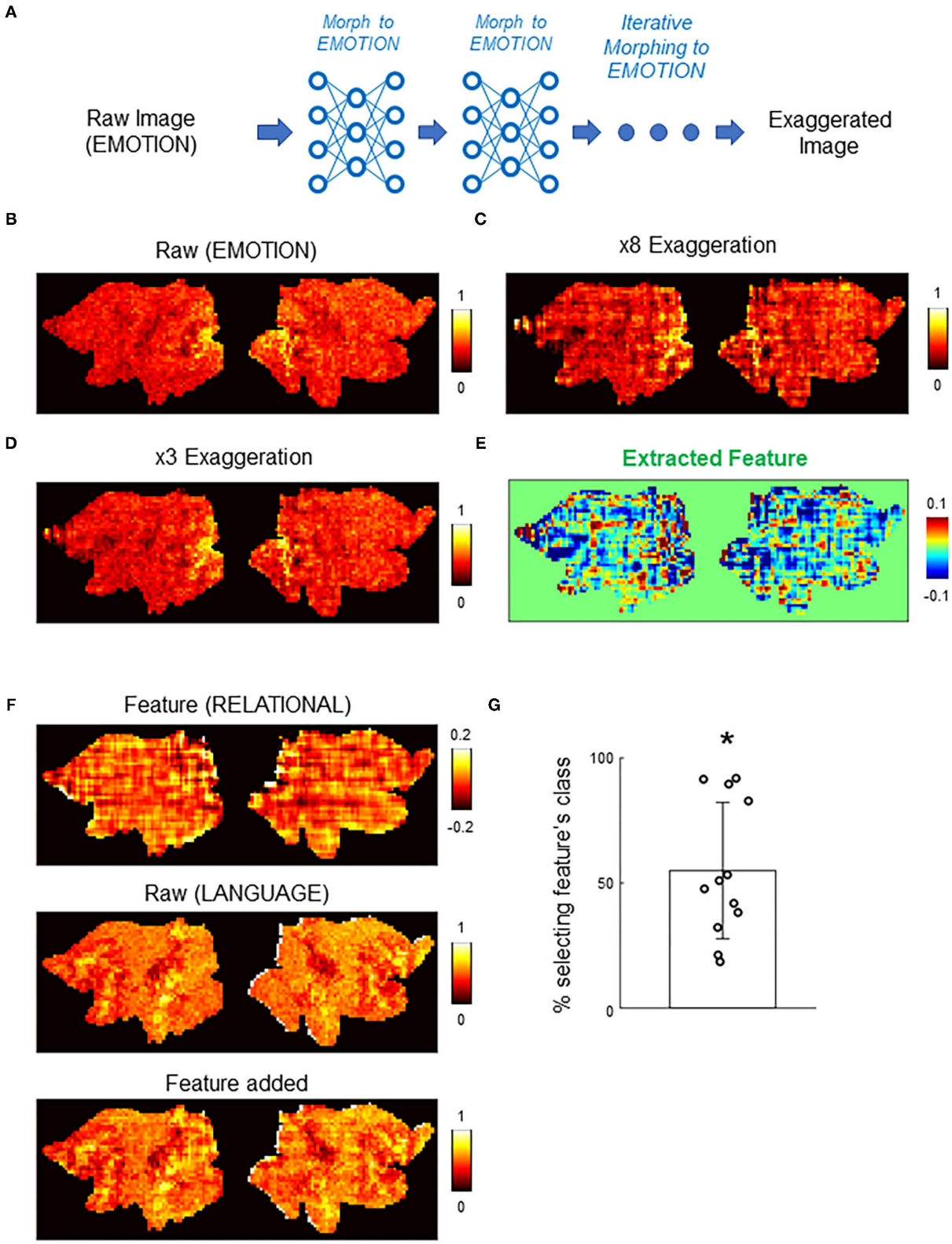


FIGURE 6 | Counterfactual exaggeration of brain activation. **(A)** Schematic of counterfactual exaggeration. A brain activation (MOTOR task in this example) was iteratively transformed toward MOTOR by CAG. This iterative transformation accentuates (exaggerates) image features that biases the classifier decision toward MOTOR. **(B–D)** Example of counterfactual exaggeration. A brain activation in the EMOTION task **(B)** was iteratively transformed toward EMOTION eight times.

(Continued)

FIGURE 6 | Images after third **(C)** and eighth **(D)** transformations are shown. **(E)** Subtle image feature enhanced by counterfactual exaggeration was isolated by taking the difference of exaggerated images. In this example, differences between exaggerated images in **(C)** and **(D)** were calculated. The resulting difference image showed a texture-like pattern. **(F)** Example of texture-like feature extracted by counterfactual exaggeration (top). Bottom panel shows the texture-like patterns added to randomly chosen raw brain activations (middle). See also **Supplementary Figure 6** for another example. **(G)** Decisions of the DNN classifier to brain activations with texture-like patterns added. Each dot represents one example texture ($N = 12$. Methods for details). Bar graph shows the mean and the standard deviation. The classifier was significantly biased toward the class of texture-like patterns (*, $p < 0.001$, Wilcoxon's sign rank test). Chance level was one of seven.

and also brain machine interface studies using DNNs [e.g., Willett et al. (2021)] can combine counterfactual explanation with saliency maps to better interpret how the patterns of brain activations are causally related to DNN decisions.

There are several limitations in this study. The training and testing of the DNN classifier and CAG were performed using only the HCP dataset. As more and more neuroimaging datasets become available to the public, researchers are starting to develop DNN classifiers trained on multiple datasets. Though it is beyond the scope of this study, explaining the DNN classifiers trained on multiple datasets would be an important future research topic. Another limitation is that this study used spatial downsampling to enable efficient learning by CAG. This was partly due to limitations in both the computational power and the dataset size. The limitation in the dataset size may be alleviated using techniques for data augmentation (Shorten and Khoshgoftaar, 2019).

It should also be emphasized that the aim of CAG is not to improve the accuracy of the DNN classifier but to provide visual explanations for the classifier's decisions. Because CAG can simultaneously take into account information from the entire brain, counterfactual explanation is different from conventional analyses of local activation patterns such as GLM and search light-based multivariate pattern analyses (Kriegeskorte et al., 2006; Jimura and Poldrack, 2012; Chikazoe et al., 2014). This characteristic of CAG is most pronounced in counterfactual exaggerations, where it discovered global texture-like patterns that could effectively bias the classifier's decisions. At present, these patterns are unlikely to reflect biologically important activity patterns. Further development of CAG and related techniques would enable the discovery of global activity patterns with biological significance beyond conventional analyses.

Conclusions

In this study, we developed CAG, a generative neural network for counterfactual brain activation that can be used to explain individual decision behaviors of DNN-based classifiers. A single CAG could handle multiple classes at the same time and learn mapping between all the pairs of classes. CAG could provide visually intuitive counterfactual explanations for a classifier's correct and incorrect decisions. These counterfactual explanations were quantitatively more effective in explaining the classifier's decision than the controls and were robust against image perturbations. Finally, beyond explaining the decision behaviors, CAG could extract subtle image features in the brain activation that were invisible to the eyes but that were exploited by the DNN classifiers. Together, these results suggest that counterfactual explanation with CAG provides a novel approach to examine and extend current neuroimaging studies using DNNs.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: S1200 release of the Human Connectome Project (http://www.humanconnectomeproject.org).

AUTHOR CONTRIBUTIONS

TM and MT conceived the study conducted analysis. KJ conducted analysis and provided reagents. TM wrote the manuscript with inputs from all authors. All authors discussed the results.

FUNDING

This study was supported by JSPS Kakenhi (20H05052 and 21H0516513 to TM, 19K20390 to TP, 19H04914 and 20K07727 to KJ, 21H02806 and 21H05060 to JC), a grant from Japan Agency for Medical Research and Development (AMED) to JC (Grant Number JP19dm0207086), a grant from Brain/MINDS Beyond (AMED) to TM and MT (Grant Number JP20dm0307031), a grant from JST-PRESTO to TM, a grant from Narishige Neuroscience Research Foundation to TM.

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.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2021.802938/full#supplementary-material

REFERENCES

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv*. 2017:arxiv:1701.07875.

Barch, D., Burgess, G., Harms, M., Petersen, S., Schlaggar, B., Corbetta, M., et al. (2013). Function in the human connectome: Task-fMRI and individual differences in behavior. *Neuroimage*. 80, 169–189. doi: 10.1016/j.neuroimage.2013.05.033

Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. (2019). "Explaining Image Classifiers by Counterfactual Generation," in *International Conference on Learning Representations (ICLR)*.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. (2018). Grad-CAM plus plus : generalized gradient-based visual explanations for deep convolutional networks. *IEEE Wint Conferen Appl Comput Vis*. 2018, 839–847. doi: 10.1109/WACV.2018.00097

Chikazoe, J., Lee, D. H., Kriegeskorte, N., and Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nat Neurosci*. 17, 1114–1122. doi: 10.1038/nn.3749

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). "StarGAN: unified generative adversarial networks for multi-domain image-to-image translation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., et al. (2019). Efficient Decision-based Black-box Adversarial Attacks on Face Recognition. *IEEE Conferen Comput Vision Pattern Recogn*. 2019, 7706–7714. doi: 10.1109/CVPR.2019.00790

Eitel, F., and Ritter, K. (2019). Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. *arXiv*. 2019:arxiv:1909.08856.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). "ImageNet-trained CNNs are biased towards textures; increasing shape bias increases robustness," in *International Conference on Learning and Representations (ICLR)*.

Glasser, M. F., Smith, S. M., Marcus, D. S., Andersson, J. L., Auerbach, E. J., Behrens, T. E., et al. (2016). The human connectome project's neuroimaging approach. *Nat. Neurosci*. 19, 1175–1187. doi: 10.1038/nn.4361

Goodkind, M. S., Sollberger, M., Gyurak, A., Rosen, H. J., Rankin, K. P., Miller, B., et al. (2012). Tracking emotional valence: the role of the orbitofrontal cortex. *Hum. Brain Mapp*. 33, 753–762. doi: 10.1002/hbm.21251

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. (2019). Counterfactual visual explanations. https://arxiv.org/list/cs/recent arXiv:1904.07451

Jimura, K., and Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*. 50, 544–552. doi: 10.1016/j.neuropsychologia.2011.11.007

Kim, T., Cha, M., Kim, H., Lee, J., Kim, J., Precup, D., et al. (2017). Learning to discover cross-domain relations with generative adversarial networks. *Int. Conf. Machine Learn*. 70, 17. doi: 10.5555/3305381.3305573

Kingma, D., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv*. 2014:arxiv:1412.6980.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U S A*. 103, 3863–3868. doi: 10.1073/pnas.0600244103

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*. 521, 436–444. doi: 10.1038/nature14539

Liu, S., Kailkhura, B., Loveland, D., and Han, Y. (2019). "Generative counterfactual introspection for explainable deep learning," in *7th Ieee Global Conference on Signal and Information Processing* (Ieee Globalsip).

Mertes, S., Huber, T., Weitz, K., Heimerl, A., and Andre, E. (2020). GANterfactual - counterfactual explanation for medical non-experts using generative adversarial learning. arXiv. 2020:arrive:2012.11905v3.

Narayanaswamy, A., Venugopalan, S., Webster, D. R., Peng, L., Corrado, G. S., Ruamviboonsuk, P., et al. (2020). Scientific discovery by generating counterfactuals using image translation. *Int. Conferen. Med. Image Comput*. 2020, 27. doi: 10.1007/978-3-030-59710-8_27

Pan, S., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*. 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Pawlowski, N., Castro, D. C., and Glocker, B. (2020). *Deep Structural Causal Models for Tractable Counterfactual Inference. Conference on Neural Information Processing Systems (NeurIPS)*.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci*. 10, 59–63. doi: 10.1016/j.tics.2005.12.004

Rolls, E. T., Cheng, W., and Feng, J. (2020). The orbitofrontal cortex: reward, emotion and depression. *Brain Commun*. 2, fcaa196. doi: 10.1093/braincomms/fcaa196

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). Grad-CAM: visual explanations from deep networks *via* gradient-based localization. *Int. J. Comput. Vision*. 128, 336–359. doi: 10.1007/s11263-019-01228-7

Shorten, C., and Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *J. Big Data*. 6,1. doi: 10.1186/s40537-019-0197-0

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*.

Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. (2020). "Explanation by progressive exaggeration," in *International Conference on Learning Representations (ICLR)*.

Tsumura, K., Kosugi, K., Hattori, Y., Aoki, R., Takeda, M., Chikazoe, J., et al. (2021). Reversible fronto-occipitotemporal signaling complements task encoding and switching under ambiguous cues. *Cereb Cortex*. 21, 11. doi: 10.1093/cercor/bhab324

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage*. 80, 62–79. doi: 10.1016/j.neuroimage.2013.05.041

Wang, P., and Vasconcelos, N. (2020). SCOUT: Self-aware discriminant counterfactual explanations. *CVPR*. 20, 8981–90. doi: 10.1109/CVPR42600.2020.00900

Wang, X., Liang, X., Jiang, Z., Nguchu, B. A., Zhou, Y., Wang, Y., et al. (2020). Decoding and mapping task states of the human brain *via* deep learning. *Hum. Brain. Mapp*. 41, 1505–1519. doi: 10.1002/hbm.24891

White, S. F., Adalio, C., Nolan, Z. T., Yang, J., Martin, A., and Blair, J. R. (2014). The amygdala's response to face and emotional information and potential category-specific modulation of temporal cortex as a function of emotion. *Front. Hum. Neurosci*. 8, 714. doi: 10.3389/fnhum.2014.00714

Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M., and Shenoy, K. V. (2021). High-performance brain-to-text communication *via* handwriting. *Nature*. 593, 249–254. doi: 10.1038/s41586-021-03506-2

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U S A*. 111, 8619–8624. doi: 10.1073/pnas.1403112111

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods*. 8, 665–670. doi: 10.1038/nmeth.1635

Zhao, Y. (2020). Fast real-time counterfactual explanations. https://arxiv.org/list/cs/recent arXiv:2007.05684

Zhu, J., Park, T., Isola, P., and Efros, A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *IEEE Int. Conferen. Comput. Vis*. 2017, 2242–2251. doi: 10.1109/ICCV.2017.244

Conflict of Interest: JC is Employed by Araya Inc.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Neuroinformatics

TECHNOLOGY AND CODE
published: 24 May 2022
doi: 10.3389/fninf.2022.862805

vol2Brain: A New Online Pipeline for Whole Brain MRI Analysis

José V. Manjón[1], José E. Romero[1], Roberto Vivo-Hernando[2], Gregorio Rubio[3], Fernando Aparici[4], Mariam de la Iglesia-Vaya[5,6] and Pierrick Coupé[7]*

[1] Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Valencia, Spain, [2] Instituto de Automática e Informática Industrial, Universitat Politècnica de València, Valencia, Spain, [3] Departamento de Matemática Aplicada, Universitat Politècnica de València, Valencia, Spain, [4] Área de Imagen Medica, Hospital Universitario y Politécnico La Fe, Valencia, Spain, [5] Unidad Mixta de Imagen Biomédica FISABIO-CIPF, Fundación Para el Fomento de la Investigación Sanitario y Biomédica de la Comunidad Valenciana, Valencia, Spain, [6] Centro de Investigación Biomédica en Red de Salud Mental, ISC III, València, Spain, [7] Centre National de la Recherche Scientifique, Univ. Bordeaux, Bordeaux INP, Laboratoire Bordelais de Recherche en Informatique, UMR5800, PICTURA, Talence, France

Automatic and reliable quantitative tools for MR brain image analysis are a very valuable resource for both clinical and research environments. In the past few years, this field has experienced many advances with successful techniques based on label fusion and more recently deep learning. However, few of them have been specifically designed to provide a dense anatomical labeling at the multiscale level and to deal with brain anatomical alterations such as white matter lesions (WML). In this work, we present a fully automatic pipeline (vol2Brain) for whole brain segmentation and analysis, which densely labels ($N > 100$) the brain while being robust to the presence of WML. This new pipeline is an evolution of our previous volBrain pipeline that extends significantly the number of regions that can be analyzed. Our proposed method is based on a fast and multiscale multi-atlas label fusion technology with systematic error correction able to provide accurate volumetric information in a few minutes. We have deployed our new pipeline within our platform volBrain (www.volbrain.upv.es), which has been already demonstrated to be an efficient and effective way to share our technology with the users worldwide.

Keywords: segmentation, brain, analysis, MRI, cloud

OPEN ACCESS

Edited by:
Fatos Tunay Yarman Vural, Middle East Technical University, Turkey

Reviewed by:
Guray Erus, University of Pennsylvania, United States
Nagesh Adluru, University of Wisconsin-Madison, United States

****Correspondence:***
José V. Manjón
jmanjon@fis.upv.es

Received: *26 January 2022*
Accepted: *07 April 2022*
Published: *24 May 2022*

Citation:
Manjón JV, Romero JE, Vivo-Hernando R, Rubio G, Aparici F, de la Iglesia-Vaya M and Coupé P (2022) vol2Brain: A New Online Pipeline for Whole Brain MRI Analysis. Front. Neuroinform. 16:862805. doi: 10.3389/fninf.2022.862805

INTRODUCTION

Quantitative brain image analysis based on MRI has become more and more popular over the last decade due to its high potential to better understand subtle changes in the normal and pathological human brain. The exponential increase in the current neuroimaging data availability and the complexity of the methods to analyze them make the development of novel approaches necessary to address challenges related to the new "Big Data" paradigm (Van Horn and Toga, 2014). Thus, automatic, robust, and reliable methods for automatic brain analysis will have a major role in the near future, most of them being powered by cost-effective cloud-based platforms.

Specifically, MRI brain structure volume estimation is being increasingly used to better understand the normal brain evolution (Coupé et al., 2017) or the progression of many neurological pathologies such as multiple sclerosis (MS, Commowick et al., 2018) or Alzheimer's disease (Coupé et al., 2019).

The quantitative estimation of the different brain structure volumes requires automatic, robust, and reliable segmentation of such structures. As manual delineation of the full brain is unfeasible

for routine brain analysis (this task is too tedious, time-consuming, and prone to reproducibility errors), many segmentation methods have been proposed over the years. Some of them were initially focused at the tissue level such as the famous Statistical Parametric mapping (SPM) (Ashburner and Friston, 2005). However, this level of detail may be insufficient to detect subtle changes in specific brain structures at early stages of the disease.

For example, hippocampus and lateral ventricle volumes can be used as early biomarkers of Alzheimer's disease. At this scale, also, cortical and subcortical gray matter (sGM) structures are of special interest for the neuroimaging community. Classic neuroimaging tools such as the well-known FSL package (Jenkinson et al., 2012) or Freesurfer (Fischl et al., 2002) have been widely used over the last 2 decades. More recently, multi-atlas label fusion segmentation techniques have been extensively applied, thanks to their ability to combine multiple atlas information minimizing mislabeling due to inaccurate registrations (Coupé et al., 2011; Wang and Yushkevich, 2013; Manjón et al., 2014; Romero et al., 2015).

However, segmentation of the whole brain into a large number of structures is still a very challenging problem even for modern multi-atlas based methods (Wang and Yushkevich, 2013; Cardoso et al., 2015; Ledig et al., 2015). The problems encountered are (1) the need of a large set of densely manually labeled brain scans and (2) the large amount of computational time needed to combine all those labeled scans to produce the final segmentation. Fortunately, a fast framework based on collaborative patch-matching was recently proposed (Giraud et al., 2016) to reduce the computational time required by multi-atlas patch-based methods.

More recently, deep leaning methods have also been proposed for brain structure segmentation. Those methods are mainly patch-based (Wachinger et al., 2018) or 2D (slice-based) (Roy et al., 2019) due to current GPU memory limitations. The current state-of-the-art whole brain deep learning methods are based on ensembles of local neural networks such as the SLANT method (Huo et al., 2019), or more recently the Assemblynet method (Coupé et al., 2020).

The aim of this study is to present a new software pipeline for whole brain analysis that we have called vol2Brain. It is based on an optimized multi-atlas label fusion scheme that has a reduced execution time, thanks to the use of our fast collaborative patch-matching approach, which has been specifically designed to deal with both normal appearing and lesioned brains (a feature that most of preceding methods ignored). This pipeline automatically provides volumetric brain information at different scales in a very simple manner through a web-based service not requiring any installation or technical requirements in a similar manner as previously done by our volBrain platform that since 2015 has processed more than 360,000 brains online worldwide. In the following sections, the new pipeline will be described, and some evidences of its quality will be presented.

MATERIALS AND METHODS

Dataset Description

In our proposed method, we used an improved version of the full Neuromorphometrics dataset (http://www.neuromorphometrics.com), which consists of 114 manually segmented brain MR volumes corresponding to subjects with ages covering almost the full lifespan (from 5 to 96 years). Dense neuroanatomical manual labeling of MRI brain scans was performed at Neuromorphometrics, Inc., following the methods described in the study by Caviness et al. (1999).

The original MRI scans were obtained from the following sources: (1) the Open Access Series of Imaging Studies (OASIS) project website (http://www.oasis-brains.org/) (N = 30), (2) the Child and Adolescent NeuroDevelopment Initiative (CANDI) Neuroimaging Access Point (http://www.nitrc.org/projects/candi_share) (N = 13), (3) the Alzheimer's Disease Neuroimaging Initiative (ADNI) project website (http://adni.loni.usc.edu/data-samples/access-data/) (N = 30), (4) the McConnell Brain Imaging Center (http://www.bic.mni.mcgill.ca/ServicesAtlases/Colin27Highres/) (N = 1), and (5) the 20Repeats dataset (http://www.oasis-brains.org/) (N = 40).

Before manual labeling, all the images were preprocessed with an automated bias field inhomogeneity correction (Arnold et al., 2001) and geometrically normalized using three anatomical landmarks [anterior commissure (AC), posterior commissure (PC), and mid-sagittal point]. The scans were reoriented and resliced so that anatomical labeling could be done in coronal planes that follow the AC-PC axis. The manual outlining was performed using an in-house software called the NVM and the exact specification of each region of interest is defined in (1) Neuromorphometrics' General Segmentation Protocol (http://neuromorphometrics.com/Seg/) and (2) the BrainCOLOR Cortical Parcellation Protocol (http://Neuromorphometrics.com/ParcellationProtocol_2010-04-05.PDF). It has to be noted that the exact protocols used to label the scans evolved over time. Because of this, not all anatomical regions were labeled in every group (label number range: max = 142, min = 136).

Dataset Correction

Right after downloading the Neuromorphometrics dataset, we performed a rigorous quality control of the dataset. We discovered that this dataset presented several issues that had to be corrected before using it.

Image Resolution, Orientation, and Size

After checking each individual file, we found that they had different acquisition orientations (coronal, sagittal, and axial). They also have different resolutions ($1 \times 1 \times 1$, $0.95 \times 0.93 \times 1.2$, $1.26 \times 1.24 \times 12$, etc.) and different volume sizes ($256 \times 256 \times 307$, $256 \times 256 \times 299$, $256 \times 256 \times 160$, etc.). To standardize them, we registered all image and corresponding label files to the MNI152 space using ANTS software, which resulted in a homogeneous dataset with axial orientation, $1 \times 1 \times 1$ mm^3 voxel resolution, and a volume size of $181 \times 217 \times 181$ voxels. We also checked the image quality and we removed 14 cases from

the original dataset that presented strong image artifacts and severe blurring effects. This resulted in a final dataset of 100 cases.

Inconsistent and Different Number of Labels

The selected 100 files from the previous step had 129 common labels from a total of 142 labels. After analyzing these 13 inconsistent labels, we decided to treat each of them in a specific manner according to the detected issue. Label file description assigns label numbers from 1 to 207. However, we found that labels 228, 229, 230, and 231 were present in some files. After checking them, we realized that labels 228 and 229 on the left corresponded to a right basal foreground (labels 75 and 76) and so we renumbered them. Labels 230 and 231 just represented few pixels in three of the cases and therefore were removed. Labels 63 and 64 (right and left vessel) were not present in all the cases (not always visible) and we decided to renumber them as a part of the putamen (labels 57 and 58), as they were located inside. We removed label 69 (optic chiasm) because it was not present in all the cases and its delineation was very inconsistent. Labels 71, 72, and 73 (cerebellar vermal lobules I-V, VI-VII, and VIII-X) were present in 74 of the 100 cases, and we decided to re-segment the inconsistent cases so that all the cases have these labels (details are given in the following section). Label 78 (corpus callosum) was only present in 25 cases, and we decided to relabel it as right and left white matter (WM, labels 44 and 45). Label 15 (5th ventricle) was very tiny and only present in a few cases (13); thus, it was relabeled as lateral ventricles (labels 51 and 52). Finally, we decided to add two new labels that we found important, i.e., external cerebrospinal fluid (CSF) (labeled as 1) and left and right WM lesions (labels 53 and 54). Details on how these labels were added are provided in the following section. After all the cleanup, the final dataset had a consistent number of 135 labels (refer to **Appendix**).

Labeling Errors

Once the dataset had a homogeneous number of labels, we inspected them to check their quality. After inspecting the dataset visually, we found that the boundaries of all the structures in sagittal and axial planes were very irregular. This is probably due to the fact that the original manual delineation was performed in the coronal plane. However, one of the main problems we found was the fact that cortical gray matter (cGM) was severely overestimated, and correspondingly, the CSF and WM were underestimated. This fact has been already highlighted by other researchers (Huo et al., 2017) who pointed out this problem in the context of cortical thickness estimation. The same problem arises in the cerebellum, although it is a bit less pronounced. To solve this problem (Huo et al., 2017), an automatic fusion of the original GM/WM maps was used, and partial volume maps were generated by the TOADS method (Bazin and Pham, 2008) to correct the cortical labels. In this study, we have followed a different approach based on the original manual segmentation and the intensity information.

First, we combined all the 135 labels into seven different classes (CSF, cGM, cerebral white matter (cWM), sGM, cerebellar gray matter (ceGM), cerebellar white matter (ceWM), and brain stem (BS)]. External CSF was not labeled in the Neuromorphometrics dataset, so we added it using volBrain (Manjón and Coupé, 2016) (we copied CSF label to those pixels that had label 0 in the original label file). Then, the median value of cGM and cWM was estimated and used to generate the partial volume maps using a linear mixing model (Manjón et al., 2008). Voxels in the cGM and cWM interface were relabeled according to their partial volume content (e.g., a cGM voxel with a cWM partial volume coefficient bigger than its corresponding cGM partial volume coefficient was relabeled as cWM). The same process was repeated for the CSF/cGM interface, the ceGM/ceWM interface, and the ceGM/CSF interface. To ensure the regularity of the new label maps, each partial volume map was regularized using a non-local means filter (Coupé et al., 2018). Finally, each case was visually revised and small labeling errors were manually corrected using the ITK-SNAP software. Most of the corrections were related with cGM in the upper part of the brain, and misclassifications of WM lesions were termed as cGM and CSF-related corrections. **Figure 1** shows an example of the cGM/cWM tissue maps before and after the correction.

After the tissue correction, the original structure labels were automatically relabeled to match the new tissue maps. Specifically, those voxels that kept the same tissue type before and after the correction kept their original labels and those that changed were automatically labeled according to the most likely label considering their position and intensity. Results were visually reviewed to assess its correctness and manually corrected when necessary. Finally, we realized that sGM structures showed important segmentation errors and we decided to re-segment them using volBrain automatic segmentation followed by manual correction when needed. **Figure 2** shows an example of the final relabeling result.

LesionBrain Dataset

One of the main goals of the proposed pipeline was to make it robust to the presence of WM lesions that normally are misclassified as gray matter (GM) in pathological brains. To this end, we included not only healthy cases but also subjects with WM lesions in our library. Specifically, 32 of the 100 cases of the previously described Neuromorphometrics dataset had WM visible lesions with a lesion load ranging from moderate to severe. We are aware that WM lesions can appear anywhere in the brain, but it is also known that they have *a priori* probability to be located in the periventricular areas among others (Coupé et al., 2018).

We found though that the number of cases with lesions on the dataset was not enough to capture the diversity of WM lesion distribution, so we decided to expand the dataset using a manually labeled MS dataset. We previously used this dataset to develop a MS segmentation method (Coupé et al., 2018).

This dataset is composed of 43 patients with MS who underwent 3T 3D-T1w MPRAGE and 3D-Fluid-Attenuated Inversion Recovery (FLAIR) MRI. We used only the T1 images, as this is the input modality of our proposed pipeline. To further increase the size of the dataset, we included the left-right flipped version of the images and labels resulting in an extended dataset of 86 cases.

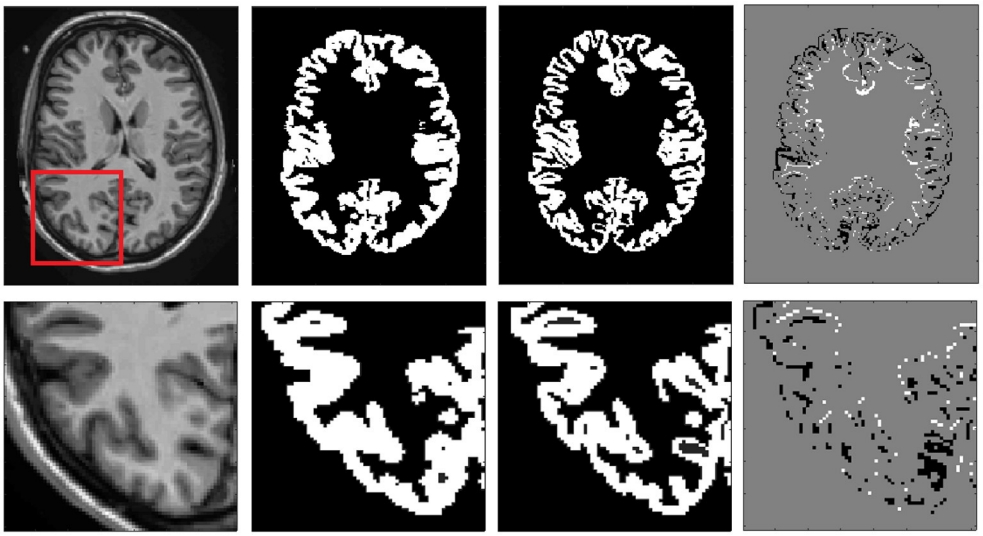


FIGURE 1 | Example of cGM tissue correction. From right to left: Reference T1 image, original cGM map, corrected cGM map, and map of changes (white means inclusion and black means removal of pf voxels). In the bottom row, a close up is shown to better highlight the differences.

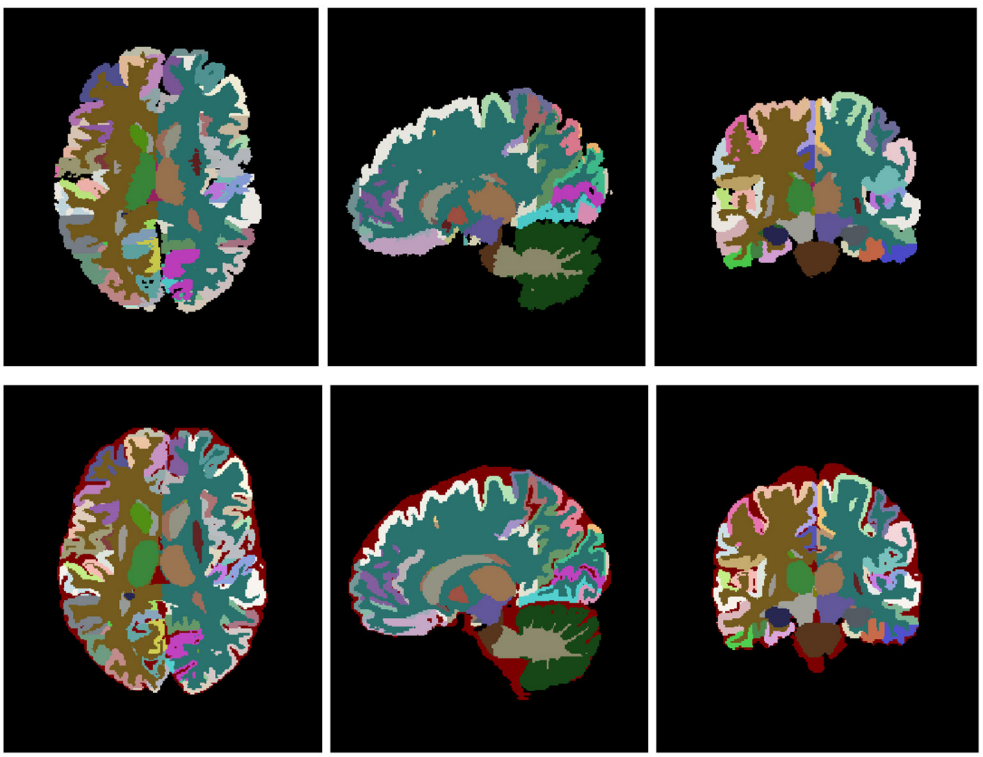


FIGURE 2 | Top row shows the original labeling and bottom row shows the corrected labeling. Note that the external CSF label has been added to the labeling protocol.

Vol2Brain Pipeline Description

The vol2Brain pipeline is a set of image processing tasks dedicated to improve the quality of the input data and to set them into a specific geometric and intensity space, to segment the different structures and to generate useful volumetric information (refer to **Figure 3** for a general overview). The vol2Brain pipeline is based on the following steps:

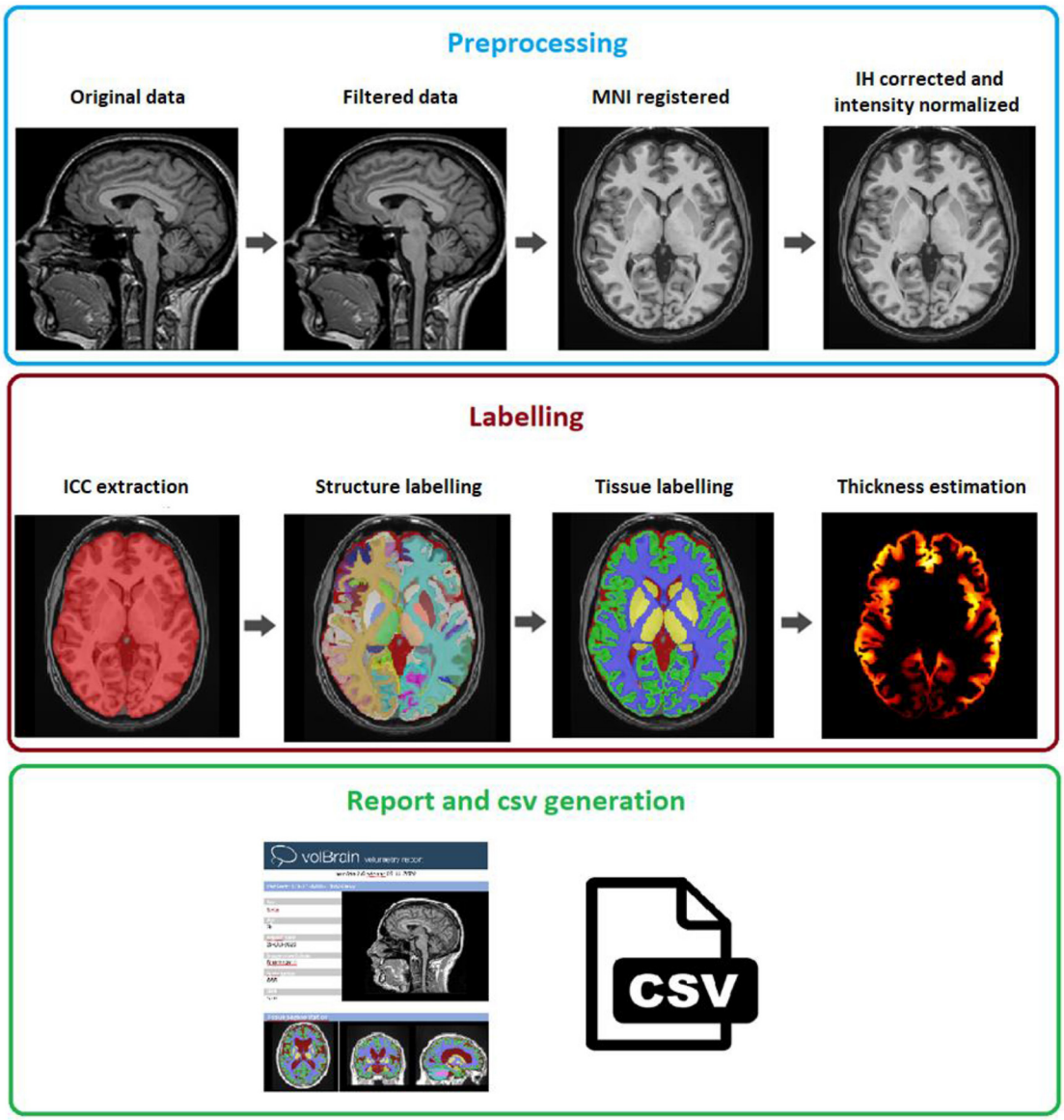


FIGURE 3 | vol2Brain pipeline scheme. In the first row, the preprocessing for any new subject is presented. In the second row, the results of the ICC extraction, structure, and tissue segmentations jointly with the cortical thickness estimation are presented. Finally, in the third row, the volumetric information is extracted and presented.

1. Preprocessing
2. Multiscale labeling and cortical thickness estimation
3. Report and csv generation

Preprocessing

We have used the same preprocessing steps as those described in the volBrain pipeline (Manjón and Coupé, 2016), as it has been demonstrated to be very robust (based in our experience processing more than 360,000 subjects worldwide). This preprocessing consists of the following steps. To improve the image quality, first, the raw image is denoised using the Spatially Adaptive Non-Local Means (SANLM) filter (Manjón et al., 2010) and inhomogeneity is corrected using the N4 method (Tustison et al., 2010). The resulting image is then affinely registered to the Montreal Neurological Institute (MNI) space using the ANTS software (Avants et al., 2008). The image in the MNI space has a size of 181 × 217 × 181 voxels with 1 mm^3 voxel resolution. Then, we used an inhomogeneity correction based on SPM8 (Ashburner and Friston, 2005) toolbox, as this model-based method has proven to be quite robust once the data are located at the MNI space. Finally, we normalized the images as per intensity by applying a piecewise linear tissue mapping based on the TMS method (Manjón et al., 2008) as described in the study by Manjón and Coupé (2016). It is worth to note that

the library images were also normalized as per intensity using the described approach so that both library and the case to be segmented share a common geometrical and intensity space.

Multiscale Labeling and Cortical Thickness Estimation

After the preprocessing, the images are ready to be segmented and measured. This segmentation is performed in several stages.

ICC Extraction

The first step in the segmentation process is the intracranial cavity extraction (ICC). This is obtained using the NICE method (Manjón et al., 2014). NICE method is based on a multi-scale non-local label fusion scheme. Details of the NICE method can be found in the study by Manjón et al. (2014). To further improve the quality of the original NICE method, we have increased the size of the original volBrain template library from 100 to 300 cases using the 100 cases of the vol2Brain library and their left-right mirrored version.

Full Brain Structure Segmentation

The dense segmentation of the full brain is based on a multiscale version of the non-local patch-based label fusion technique (Coupé et al., 2011) wherein patches of the subject to be segmented are compared with patches of the training library to look for similar patterns within a predefined search volume to assign the proper label v as can be seen in the following equation:

$$v(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \in V_i} w(x_i, x_{s,j}) y_{s,j}}{\sum_{s=1}^{N} \sum_{j \in V_i} w(x_i, x_{s,j})} \quad (1)$$

where V_i corresponds to the search volume, N is the number of subjects in the templates library, $y_{s,j}$ is the label of the voxel $x_{s,j}$ at the position j in the library subject s, and $w(x_i, x_{s,j})$ is the patch similarity defined as:

$$w\left(x_i, x_{s,j}\right) = \exp^{\frac{-D_{i,j,s}}{h^2}} \quad (2)$$

$$D_{i,j,s} = \left\| P(x_i) - P(x_{s,j}) \right\|_2^2 \quad (3)$$

where $P(x_i)$ is the patch centered at x_i, $P(x_{s,j})$ is the patch centered at x_j in the templates, and $\|.\|_2$ is the normalized L2 norm (normalized by the number of elements) calculated from the distance between each pair of voxels from both patches $P(x_i)$ and $P(x_{s,j})$. h is a normalization parameter that is estimated from the minimum of all patch distances within the search volume.

However, exhaustive patch comparison process is very time-consuming (even in reduced neighborhoods, i.e., when the search volume V is small). To reduce the computational burden of this process, we have used a multiscale adaptation of the OPAL method (Giraud et al., 2016) previously proposed in the study by Romero et al. (2017), which takes benefit from the concept of Approximate Nearest Neighbor Fields (ANNF). To further speed up the process, we processed only those voxels that were segmented as ICC by the NICE method.

In patch-based segmentation, the patch size is a key parameter that is strongly related to the structure to be segmented and image resolution. It can be seen in the literature that multi-scale approaches improve segmentation results (Manjón et al., 2014). In the OPAL method (Giraud et al., 2016), independent and simultaneous multi-scale and multi-feature artificial neural networks (ANN) fields were computed. Thus, we have followed a multi-scale approach in which several different ANNs are computed for different patch sizes resulting in different label probability maps that have to be combined. In this study, two patch sizes are used, and an adaptive weighting scheme is proposed to fuse these maps (Equation 3).

$$p(l) = \alpha\, p_1(l) + (1 - \alpha) p_2(l) \quad (4)$$

where $p_1(l)$ is the probability map of patch-size $3 \times 3 \times 3$ volxels for label l, $p_2(l)$ is the probability map of patch-size $5 \times 5 \times 5$ voxels for label l, $p(l)$ is the final probability map for label l, and $\alpha \in [0,1]$ is the probability mixing coefficient.

Systematic Error Correction

Any segmentation method is subject to both random and systematic errors. The first error type can be typically minimized by using bootstrapped estimations. Fortunately, the non-local label fusion technique estimates the voxel label averaging the votes of many patches, which naturally reduces the random classification error. Unfortunately, systematic errors cannot be reduced using this strategy, as they are not random. However, due to its nature, this systematic bias can be learned, and later, this knowledge can be used to correct the segmentation output (Wang and Yushkevich, 2013).

In the study by Romero et al. (2017), we proposed an error corrector method based on a patch-based ensemble of neural networks (PEC for Patch-based Ensemble Corrector) to increase the segmentation accuracy by reducing the systematic errors. Specifically, a shallow neural network ensemble is trained with image patches of sizes $3 \times 3 \times 3$ voxels (fully sampled) and $7 \times 7 \times 7$ voxels (subsampled by skipping two voxels at each dimension) from the T1w images, the automatic segmentations, a distance map value, and their x, y, and z coordinates at MNI152 space. The distance map we used is calculated for the whole structure as the distance in voxels to the structure contour. This results in a vector of 112 features that are mapped to a patch of manual segmentations of size $3 \times 3 \times 3$ voxels. We used a multilayer perceptron with two hidden layers of size 83 and 55 neurons resulting in a network with a topology of $112 \times 83 \times 55 \times 27$ neurons. An ensemble of 10 neural networks was trained using a boosting strategy. Each new network was trained with a different subset of data, which was selected by giving a higher probability of selection to those samples that were misclassified in the previous ensemble. More details can be found in the original study (Romero et al., 2017).

Multiscale Label Generation

Once the full brain segmentation is performed, different scale versions were computed by combining several labels to generate more generic ones and allowing a multiscale brain analysis. The 135 labels were combined to create a tissue-type segmentation map, including eight different tissues [CSF, cGM, cWM, sGM,

ceGM, ceWM, BS, and white matter lesions (WML)]. The cGM and cWM maps will be later used to compute the cortical thickness. Also, cerebrum lobe maps were created by combining cortical GM structures. These maps will be used later to compute the lobe-specific volumes and thickness.

Cortical Thickness Estimation

To estimate the cGM thickness, we have used the DiReCT method. DiReCT was introduced in the study by Das et al. (2009) and was made available in ANTs under the program named *KellyKapowski*. This method is based on the use of a dense non-linear registration to estimate the distance between the inner and the outer parts of the cGM. Cortical thickness per cortical label and per lobe were estimated from the thickness map and the corresponding segmentation maps (Tustison et al., 2014).

Report Generation

The output produced by the vol2Brain pipeline consists in a pdf and csv files. These files summarize the volumes and asymmetry ratios estimated from the images. If the user provides sex and age of the submitted subject, population-based normal volumes and asymmetry bounds for all structures are added for reference purposes. These normality bounds were automatically estimated from the IXI dataset (https://brain-development.org/ixi-dataset/), which contains almost 600 normal subjects covering most of the adult lifespan. We are aware that one of the most important sources of variability is the use of different scanners to build the normative values (although the use of our preprocessing reduces this variability). In the near future, we will extend the dataset to have a larger and more representative sample of the population as we already did for the volBrain pipeline (Coupé et al., 2017).

Furthermore, the user can access to its user area through volBrain website to download the resulting nifti files containing the segmentations at different scales (both in native and MNI space). An example of the volumetric report produced by vol2Brain is shown in **Appendix**.

EXPERIMENTS AND RESULTS

In this section, some experimental results are shown to highlight the accuracy and reproducibility of the proposed pipeline. A leave-two-out procedure was performed for the 100 subjects of the library (i.e., excluding the case to be segmented and its mirrored version). In the dataset, there are 19 cases that were scanned and labeled twice for the purpose of reproducibility estimation. In this case, a leave-four-out procedure was applied to avoid any problem (i.e., excluding the case to be segmented and its mirrored version of the two acquisitions of the same subject). To measure the segmentation quality, the dice index (Zijdenbos et al., 1994) was computed by comparing the manual segmentations with the segmentations obtained with our method. A visual example of the automatic segmentation results is shown in **Figure 4**.

Results

Since presenting dice results of the 135 labels would be impractical, we have decided to show the average results for cortical and non-cortical labels as done in previous studies (Wang and Yushkevich, 2013). In **Table 1**, the results of the proposed method are shown with and without the corrective learning step (PEC) to show the impact that this postprocessing has in the final results (it improved the results in all the cases).

To further explore the results, we separated them by dataset, as it is well-known that results within the same dataset are normally better than across the datasets. This allows to explore the generalization capabilities of the proposed method. Results are summarized in **Table 2**. As can be seen, results of the OASIS dataset were the best among the datasets. This makes perfect sense, as precisely, this dataset is the largest. CANDI dataset showed the worst results. This dataset had the worst image quality, which somehow explains these results.

One of the objectives of the proposed method was to be able to deal with images with white mater lesions. This is fundamental, as if we do not take into account those regions, they are normally misclassified as a cGM or sGM (which also affects the cortical thickness estimation) (Dadar et al., 2021). The results of WM lesion segmentation are summarized in **Table 3** (left and right lesions were considered together). We separated the results by lesion volume, as it is well-known that small lesions are more difficult to segment than the big ones (Manjón et al., 2018).

Once the full brain is segmented into 135 labels, those labels are grouped together to provide information at different anatomical scales. Specifically, eight different tissue labels are generated. Dice results are summarized in **Table 4**.

Method Reproducibility

A very important feature for a measurement method is its reproducibility. To measure the reproducibility of the proposed method, we used a subset of our library. Specifically, we used 19 cases of the OASIS subset that were scanned and labeled twice. In this case, we have two sources of variability, which are related to the inter-image changes and manual labeling differences. To measure the reproducibility, we computed the dice coefficient between the two different segmentations (of each case and its repetition). This was done for both the manual segmentation (that we used as a reference) and the automatic one. Results are summarized in **Table 5**. As can be seen, the proposed method showed a slightly superior reproducibility than manual labeling.

Method Comparison

It is difficult to compare the proposed method with similar state-of-the-art methods such as Freesurfer, as the labeling protocol is slightly different. For this reason, we have used as a freely available and well-known method called Joint Label Fusion as a reference (Wang and Yushkevich, 2013). This method is a state-of-the-art multi-atlas segmentation approach. To make it fully comparable, we used the corrected cases of our library as the atlas library. We summarized the results of the comparison in **Table 6**. We compared our proposed method with two versions of the JLF approach, one using an affine registered library (linear) and another using a non-linear registered library. It is worth to

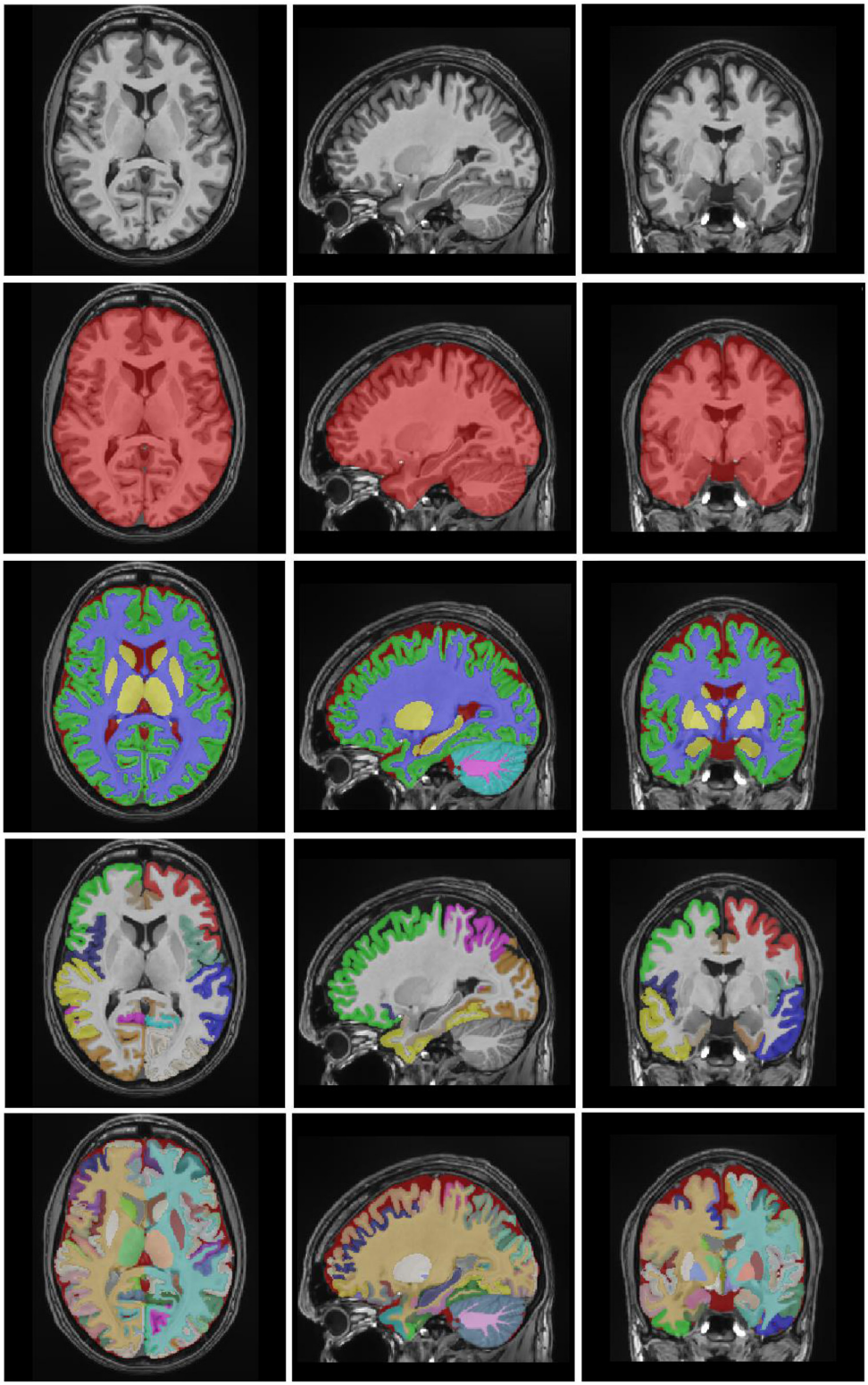


FIGURE 4 | Example results of vol2Brain. T1 image, ICC mask, brain tissues, lobes, and structures.

TABLE 1 | Proposed method dice results.

Method	All labels	Cortical labels	Non-cortical labels
Our method	0.8190 ± 0.0300	0.7912 ± 0.0397	0.8929 ± 0.0173
Our method + PEC	**0.8262 ± 0.0257**	**0.7996 ± 0.0347**	**0.8969 ± 0.0157**

*The mean dice is evaluated on all the considered labels (135 without background). *Best results highligthed in bold.*

TABLE 2 | Proposed method overall dice results for the full dataset and for each of the subsets.

All (*N* = 100)	OASIS (*N* = 68)	CANDI (*N* = 6)	ADNI (*N* = 25)	COLIN (*N* = 1)
0.8262 ± 0.0257	0.8353 ± 0.0233	0.7831± 0.0326	0.8111 ± 0.0142	0.8353

TABLE 3 | Proposed method lesion dice results.

Method	Small (*N* = 76)	Medium (*N* = 21)	Big (*N* = 3)	Avg (*N* = 100)
Lesion	0.5767 ± 0.1486	0.8281 ± 0.0500	0.8467 ± 0.0524	0.6440 ± 0.1589

**Small (<4 ml), Medium (4–18 ml), Big (>18 ml).*

TABLE 4 | Proposed method dice results for each brain tissue.

CSF	cGM	cWM	sGM
0.9006 ± 0.0307	0.9543 ± 0.0144	0.9669 ± 0.0131	0.9518 ± 0.0114
ceGM	**ceWM**	**BS**	**Lesion**
0.9644 ± 0.0172	0.9448 ± 0.0363	0.9693 ± 0.0137	0.6440 ± 0.1589

TABLE 5 | Proposed method dice results.

Method	All labels	Cortical labels	Non-cortical labels
vol2Brain	0.8405 ± 0.0181	0.8234 ± 0.0206	0.8856 ± 0.0158
Manual	0.8368 ± 0.0171	0.8198 ± 0.0200	0.8818 ± 0.0163

The mean dice is evaluated on all the considered labels (135 without background).

TABLE 6 | Proposed method dice results compared with the results of two versions of JLF method.

Method	All labels	Cortical labels	Non-cortical labels
vol2Brain	0.8262 ± 0.0257	0.7996 ± 0.0347	0.8969 ± 0.0157
JLF (linear)	0.7369 ± 0.0292	0.7016 ± 0.0337	0.8305 ± 0.0241
JLF (non-linear)	0.7591 ± 0.0252	0.7327 ± 0.0288	0.8291 ± 0.0228

note the proposed method uses only a linearly registered library (i.e., no non-linear registration was used). As can be noticed, the proposed method was far superior to both versions.

Computational Time

The proposed method takes around 20 min on average to complete the whole pipeline (including cortical thickness estimation and report generation). JLF method takes around only 2 h for structure segmentations without cortical thickness estimation (excluding the preprocessing, which includes several hours of non-linear registration depending on the number of atlases used). Freesurfer normally takes around 6 h to perform the complete analysis (which also includes surface extraction).

DISCUSSION

We have presented a new pipeline for full brain segmentation (vol2Brain) that is able to segment the brain into 135 different regions in a very efficient and accurate manner. The proposed method also integrates these 135 regions to provide measures at different anatomical scales, including brain tissues and lobes. It also provides cortical thickness measurements per cortical structure and lobe displayed into an automatic report summarizing the results (refer to **Appendix**).

To create vol2Brain pipeline, we had to create a template library that integrates all the anatomical information needed to perform the labeling process. This was a long and laborious work, as the original library obtained from Neuromorphometrics did not meet the required quality and we had to invest a significant amount of time to make it ready to use. To create this library, we homogenized the image resolution, orientation, and size of the images, removed and relabeled inconsistent labels, and corrected systematic labeling errors. Besides, we extended the labeling protocol by adding external CSF and WM lesions. As a result, we generated a highly consistent and high-quality library that not only allowed to develop the current proposed pipeline but will also be a valuable resource for future developments.

The proposed method is based on patch-based multi-atlas label fusion technology. Specifically, we have used an optimized version of non-local label fusion called OPAL that efficiently finds patch matches needed to label each voxel in the brain by reducing the required time to label the full brain from hours to minutes. To further improve the results, we have used a patch-based error corrector, which has been previously used in other segmentation problems such as hippocampus subfield labeling (Romero et al., 2017) or cerebellum lobules (Carass et al., 2018).

We measured the results of the proposed pipeline using a LOO methodology and achieved an average dice value of 0.8262. This result was obtained from four different sub-datasets ranking from 0.7831 to 0.8353 showing a good generalization of the proposed method. This result was quite close to the manual intraobserver accuracy that was estimated as 0.8363 using a reduced dataset. We also compared the proposed method with a related currently available state-of-the-art method for full brain labeling. We demonstrated that vol2Brain was not only far superior to the linear (0.8262 vs. 0.7369) and nonlinear (0.8262 vs. 0.7591) versions of JLF method but also more efficient with a temporal cost of minutes compared with hours.

The proposed vol2Brain pipeline is already available through our volBrain platform (https://volbrain.upv.es). As compared to the rest of the volBrain platform pipelines, this pipeline receives an anonymized and compressed nifti file (a T1-weighted image in the case of vol2Brain) through the website and reports the results 20 min later by sending an email to the user. The user can also download the segmentation nifti files through the user

area of volBrain platform (an example of the pdf report is shown in **Appendix**).

We hope that the accuracy and efficiency of the proposed method and the ease of use through the volBrain platform will boost the anatomical analysis of the normal and pathological brain (especially on those cases with WM lesions).

CONCLUSION

In this study, we present a novel pipeline to densely segment the brain and to provide measurements of different features at different anatomical scales in an accurate and efficient manner. The proposed pipeline has been compared with a state-of-the-art-related method showing competitive results in terms of accuracy and computational time. Finally, we hope that the online accessibility of the proposed pipeline will facilitate the access of any user around the world to the proposed pipeline making their MRI data analysis simpler and more efficient.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not publicly available because the dataset is currently protected by a license. Requests to access the datasets should be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JM and PC designed and implemented the software. JR helped in the experiments and coding. RV-H, GR, MI-V, and FA helped in the library definition and report generation. All authors writed and reviewed the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Spanish DPI2017-87743-R grant from the Ministerio de Economia, Industria y Competitividad of Spain. This work was benefited from the support of the project DeepvolBrain of the French National Research Agency (ANR-18-CE45-0013). This study was achieved within the context of the Laboratory of Excellence TRAIL ANR-10-LABX-57 for the BigDataBrain project.

ACKNOWLEDGMENTS

We thank the Investments for the future Program IdEx Bordeaux (ANR-10-IDEX- 03- 02, HL-MRI Project), Cluster of excellence CPU, and the CNRS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2022.862805/full#supplementary-material

REFERENCES

Arnold, J. B., Liow, J. S., Schaper, K. A., Stern, J. J., Sled, J. G., Shattuck, D. W., et al. (2001). Qualitative and quantitative evaluation of six algorithms for correcting intensity nonuniformity effects. *Neuroimage* 13, 931–943. doi: 10.1006/nimg.2001.0756

Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *Neuroimage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41 doi: 10.1016/j.media.2007.06.004

Bazin, P. L., and Pham, D. L. (2008). Homeomorphic brain image segmentation with topological and statistical atlases. *Med. Image Anal.* 12, 616–625. doi: 10.1016/j.media.2008.06.008

Carass, A., Cuzzocreo, J. L., Han, S., Hernandez-Castillo, C. R., Rasser, P. E., Ganz, M., et al. (2018). Comparing fully automated state-of-the-art cerebellum parcellation from Magnetic Resonance Imaging. *Neuroimage* 183, 150–172. doi: 10.1016/j.neuroimage.2018.08.003

Cardoso, M. J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., et al. (2015). geodesic information flows: spatially-variant graphs and their application to segmentation and fusion. *IEEE Trans. Med. Imaging* 34, 1976–1988. doi: 10.1109/TMI.2015.2418298

Caviness, V. S., Lange, N. T., Makris, N., Herbert, M. R., and Kennedy, D. N. (1999). MRI based brain volumetrics: emergence of a developmental brain science. *Brain Dev.* 21, 289–295. doi: 10.1016/S0387-7604(99)00022-4

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8, 13650. doi: 10.1038/s41598-018-31911-7

Coupé P., Catheline, G., Lanuza, E., and Manjón, J. V. (2017). Towards a unified analysis of brain maturation and aging across the entire lifespan: a MRI analysis. *Human Brain Mapping* 38, 5501–5518. doi: 10.1002/hbm.23743

Coupé P., Manjón, J. V., Fonov, V., Pruessner, J., Robles, M., and Collins, D. L. (2011). Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. *Neuroimage* 54, 940–954. doi: 10.1016/j.neuroimage.2010.09.018

Coupé P., Manjón, J. V., Lanuza, E., and Catheline, G. (2019). Timeline of brain alterations in Alzheimer's disease across the entire lifespan. *Sci. Rep.* 9, 3998. doi: 10.1038/s41598-019-39809-8

Coupé P., Mansencal, B., Clément, M., Giraud, R., Denis de Senneville, B., Ta, V. T., et al. (2020). AssemblyNet: a large ensemble of CNNs for 3D whole brain MRI segmentation. *Neuroimage* 219, 117026. doi: 10.1016/j.neuroimage.2020.117026

Coupé P., Tourdias, T., Linck, P., Romero, J. E., and Manjon, J. V. (2018). *LesionBrain: An Online Tool for White Matter Lesion, Segmentation. PatchMI workshop, MICCA2018* (Granada). doi: 10.1007/978-3-030-00500-9_11

Dadar, M., Potvin, O., Camicioli, R., and Duchesne, S. (2021). Beware of white matter hyperintensities causing systematic errors in FreeSurfer gray matter segmentations!. *Hum. Brain Mapp.* 42, 2734–2745. doi: 10.1002/hbm.25398

Das, S. R., Avants, B. B., Grossman, M., and Gee, J. C. (2009). Registration based cortical thickness measurement. *Neuroimage* 45, 867–879. doi: 10.1016/j.neuroimage.2008.12.016

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355. doi: 10.1016/S0896-6273(02)00569-X

Giraud, R., Ta, V. T., Papadakis, N., Manjón, J. V., Collins, D. L., and Coupé, P. (2016). An optimized PatchMatch for multi-scale and multi-feature label fusion. *Neuroimage* 124, 770–782. doi: 10.1016/j.neuroimage.2015.07.076

Huo, Y., Plassard, A. J., Carass, A., Resnick, S. M., Pham, D. L., Prince, J. L., et al. (2017). Consistent cortical reconstruction and multi-atlas brain segmentation. *Neuroimage*138, 197–210. doi: 10.1016/j.neuroimage.2016.05.030

Huo, Y., Xu, Z., Xiong, Y., Aboud, K., Parvathaneni, P., Bao, S., et al. (2019). 3D whole brain segmentation using spatially localized atlas network tiles. *Neuroimage* 194, 105–119. doi: 10.1016/j.neuroimage.2019.03.041

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Ledig, C., Heckemann, R. A., Hammers, A., Lopez, J. C., Newcombe, V., Makropoulos, A., et al. (2015). Robust whole-brain segmentation: application to traumatic brain injury. *Med. Image Anal.* 21, 40–58 doi: 10.1016/j.media.2014.12.003

Manjón, J. V., and Coupé P. (2016). volBrain: an online MRI brain volumetry system. *Front. Neuroinform.* 10, 1–30. doi: 10.3389/fninf.2016.00030

Manjón, J. V., Coupé, P., Martí-Bonmatí, L., Robles, M., and Collins, L. (2010). Adaptive non-local means denoising of MR images with spatially varying noise levels. *J. Magn. Reson. Imaging* 31, 192–203. doi: 10.1002/jmri.22003

Manjón, J. V., Coupé, P., Raniga, P., Xi, Y., Desmond, P., Fripp, J., et al. (2018). MRI white matter lesion segmentation using an ensemble of neural networks and overcomplete patch-based voting. *Comput. Med. Imaging Graphics* 69, 43–51. doi: 10.1016/j.compmedimag.2018.05.001

Manjón, J. V., Eskildsen, S. F., Coupé, P., Romero, J. E., Collins, D. L., and Robles, M. (2014). Non-local intracranial cavity extraction. *IJBI* 2014, 820205. doi: 10.1155/2014/820205

Manjón, J. V., Tohka, J., García-Martí, G., Carbonell-Caballero, J., Lull, J. J., Martí-Bonmatí, L., et al. (2008). Robust MRI brain tissue parameter estimation by multistage outlier rejection. *Magn. Reson. Med.* 59, 866–873. doi: 10.1002/mrm.21521

Romero, J. E., Coupé, P., Giraud, R., Ta, V. T., Fonov, V., Park, M., et al. (2017). CERES: a new cerebellum lobule segmentation method. *Neuroimage* 147, 916–924. doi: 10.1016/j.neuroimage.2016.11.003

Romero, J. E., Manjón, J. V., Tohka, J., Coupé, P., and Robles, M. (2015). Non-local automatic brain hemisphere segmentation. *Magn. Reson. Imaging* 33, 474–484. doi: 10.1016/j.mri.2015.02.005

Roy, A. G., Conjeti, S., Navab, N., and Wachinger, C. (2019). QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. *Neuroimage* 186, 713–727. doi: 10.1016/j.neuroimage.2018.11.042

Tustison, N., Cook, P., Klein, A., Song, G., Das, S. R., Duda, J. T., et al. (2014). Large-scale evaluation of ANTs and FreeSurfer cortical thickness measurements. *Neuroimage* 1, 166–179. doi: 10.1016/j.neuroimage.2014.05.044

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., and Yushkevich, P. A. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Van Horn, J. D., and Toga, A. W. (2014). Human neuroimaging as a "Big Data" science. *Brain Imaging Behav.* 8, 323–331. doi: 10.1007/s11682-013-9255-y

Wachinger, C., Reuter, M., and Klein, T. (2018). DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. *Neuroimage* 170, 434–445 doi: 10.1016/j.neuroimage.2017.02.035

Wang, H., and Yushkevich, P. (2013). Multi-atlas segmentation with joint label fusion and corrective learning—an open source implementation. *Front. Neuroinform.* 7, 27. doi: 10.3389/fninf.2013.00027

Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., and Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans. Med. Imaging* 13, 716–724. doi: 10.1109/42.363096

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Neuroinformatics

METHODS
published: 23 June 2022
doi: 10.3389/fninf.2022.835772

PyMVPD: A Toolbox for Multivariate Pattern Dependence

*Mengting Fang, Craig Poskanzer and Stefano Anzellotti**

Department of Psychology and Neuroscience, Boston College, Boston, MA, United States

Cognitive tasks engage multiple brain regions. Studying how these regions interact is key to understand the neural bases of cognition. Standard approaches to model the interactions between brain regions rely on univariate statistical dependence. However, newly developed methods can capture multivariate dependence. Multivariate pattern dependence (MVPD) is a powerful and flexible approach that trains and tests multivariate models of the interactions between brain regions using independent data. In this article, we introduce PyMVPD: an open source toolbox for multivariate pattern dependence. The toolbox includes linear regression models and artificial neural network models of the interactions between regions. It is designed to be easily customizable. We demonstrate example applications of PyMVPD using well-studied seed regions such as the fusiform face area (FFA) and the parahippocampal place area (PPA). Next, we compare the performance of different model architectures. Overall, artificial neural networks outperform linear regression. Importantly, the best performing architecture is region-dependent: MVPD subdivides cortex in distinct, contiguous regions whose interaction with FFA and PPA is best captured by different models.

Keywords: multivariate pattern dependence, connectivity, fMRI, deep networks, toolbox

Edited by:
Itir Onal Ertugrul,
Utrecht University, Netherlands

Reviewed by:
Kai M. Görgen,
Charite University Hospital Berlin, Germany
Giancarlo Valente,
Maastricht University, Netherlands

***Correspondence:**
Stefano Anzellotti
stefano.anzellotti@bc.edu

Received: *14 December 2021*
Accepted: *17 May 2022*
Published: *23 June 2022*

Citation:
Fang M, Poskanzer C and Anzellotti S (2022) PyMVPD: A Toolbox for Multivariate Pattern Dependence. Front. Neuroinform. 16:835772. doi: 10.3389/fninf.2022.835772

1. INTRODUCTION

Cognitive processes recruit multiple brain regions. Understanding which of these regions interact, and what computations are performed by their interactions, remains a fundamental question in cognitive neuroscience. In an effort to answer this question, a large literature has used measures of the statistical dependence between functional responses in different brain regions. The most widespread approach adopted in this literature—"functional connectivity"—computes the correlation between the timecourses of responses in different brain regions, and has been applied to both resting state fMRI and task-based fMRI (Horwitz et al., 1992; Friston, 1994; Greicius et al., 2003; Schaefer et al., 2018). Other approaches, such as Granger Causality (Granger, 1969; Goebel et al., 2003) and Dynamic Causal Modeling (Friston et al., 2003), have been developed to investigate the directionality of interactions.

In a separate literature, researchers studying the content of neural representations have developed techniques that leverage the multivariate structure of activity patterns (multivariate pattern analysis—MVPA) to decode information from fMRI data (Norman et al., 2006), and to study the similarity between the responses to different stimuli (Kriegeskorte et al., 2008). The success of MVPA has inspired the development of multivariate approaches to study the statistical dependence between brain regions (Anzellotti and Coutanche, 2018; Basti et al., 2020).

One approach, "Informational Connectivity," computes the trial-by-trial decoding accuracy for a given categorization in multiple regions, and correlates the decoding accuracy achieved with data from one region with the accuracy achieved with the other region across trials (Coutanche and Thompson-Schill, 2013). Another approach uses multivariate distance correlation to capture the statistical dependence between regions (Geerligs et al., 2016)—thanks to this strategy, it can also be applied to resting state studies, in which different conditions that can be categorized are not available.

Among multivariate approaches to study the interactions between brain regions, multivariate pattern dependence (MVPD, Anzellotti et al., 2017a; Li et al., 2019) is unique in that it trains and tests models of the interactions between brain regions using independent subsets of data, evaluating out-of-sample generalization. Like multivariate distance correlation, MVPD can be applied to both task data and resting state data. Additionally, MVPD can flexibly use a variety of models of dependence, with the potential to incorporate regularization, and to capture linear, as well as non-linear, interactions between brain regions. Of course, the use of holdout data for model evaluation has been previously adopted for applications outside the field of connectivity—indeed, it is also used in MVPA (Haxby et al., 2001; Haynes and Rees, 2006), and it has an even longer history in machine learning (see for instance Lachenbruch and Mickey, 1968). Similarly, the use of multivariate methods is also present in MVPA, and has a longer history in Science (Pearson, 1903).

Given the complex nature of the MVPD, a dedicated toolbox can provide researchers with a more accessible entry point to adopt this method. Several toolboxes have been developed for MVPA (Hanke et al., 2009; Hebart et al., 2015; Oosterhof et al., 2016; Treder, 2020), these toolboxes played an important role for the diffusion of MVPA analyses (as evidenced by the many times they have been cited). Here, we introduce a freely available open-source toolbox for MVPD, developed in Python: PyMVPD. The toolbox offers a set of functions for performing MVPD analyses, organized around a simple workflow. It also includes example Python scripts for several MVPD models, including linear regression models that were used in previous MVPD publications (Anzellotti et al., 2017a; Li et al., 2019), and new models based on artificial neural networks. The models are accompanied by algorithms that can be used to evaluate their performance. PyMVPD scripts have been designed so that they can be easily customized, enabling users to expand the toolbox to address their needs.

The full PyMVPD toolbox (including the artificial neural network models) requires a working installation of PyTorch. The use of CUDA and general purpose graphics processing units (GPGPUs) is recommended. For users who might not need artificial neural networks, we also make available a lite version of the toolbox, that does not require PyTorch. Both versions of PyMVPD can be installed with PyPI.

In the remainder of the article, a brief technical introduction to MVPD is followed by a description of PyMVPD implementation and the analysis workflow (**Figure 1**). Next, the algorithms are validated by analyzing a publicly available dataset—the StudyForrest dataset (Sengupta et al., 2016). Finally, the performance of different types of models is assessed, comparing the predictive accuracy of linear regression and artificial neural networks.

2. METHODS

2.1. MVPD

Multivariate pattern dependence (MVPD) is a novel technique that analyzes the statistical dependence between brain regions in terms of the multivariate relationship between their patterns of response. Compared with traditional methods used for connectivity analysis, MVPD has two main advantages (Anzellotti et al., 2017a). First, MVPD preserves the fine-grained information that can be lost by spatially averaging in mean-based univariate connectivity. By doing so, MVPD improves sensitivity as compared to univariate methods such as standard functional connectivity (Anzellotti et al., 2017a). This choice is motivated by the success of multivariate analysis methods developed outside the field of connectivity (MVPA, Haxby et al., 2001; Haynes and Rees, 2006; Norman et al., 2006). Second, MVPD is trained and tested with independent subsets of data. As a consequence, it is resilient to overfitting: in MVPD it is not sufficient for a model of the interactions between two regions to provide a good fit for a set of data, the model also has to generate accurate predictions for a separate set of data that was not used to tune the model's parameters (the "testing" data). This is a key feature of MVPD: it guarantees a more stringent test of the interactions between brain regions. Note that however this procedure does not remove the need for denoising methods: some sources of noise can produce shared effects across multiple regions. A previous study investigated the effectiveness of different denoising techniques for MVPD (Li et al., 2019); among the techniques tested, CompCor (Behzadi et al., 2007) was the most effective, therefore we used CompCor for denoising in this study.

Due to the use of separate sets of data for training and testing, MVPD benefits from fMRI datasets that include multiple experimental runs within each participant. This way, there is a sufficient amount of data to train the models, even after holding some out for testing. The amount of training data within a participant affects the model's ability to generate accurate predictions for the testing data. For this reason, datasets that include a very short amount of data within each participant (e.g., a 5-min resting state scan) are not well-suited for this type of analysis—in this respect, MVPD is similar to multivariate pattern analysis (MVPA).

The number of participants needed for MVPD analysis might vary depending on the brain regions that are being investigated. In previous studies, numbers of participants similar to the ones used for MVPA have produced robust results (Anzellotti et al., 2017a; Li et al., 2019). Based on these considerations, in the present work we used the StudyForrest dataset (Hanke et al., 2016), a publicly available dataset that has been used in several MVPA studies. As compared to large datasets used in functional connectivity (such as the Human Connectome Project dataset, Smith et al., 2013), the number of participants in the StudyForrest dataset is relatively small (14 subjects for analysis),

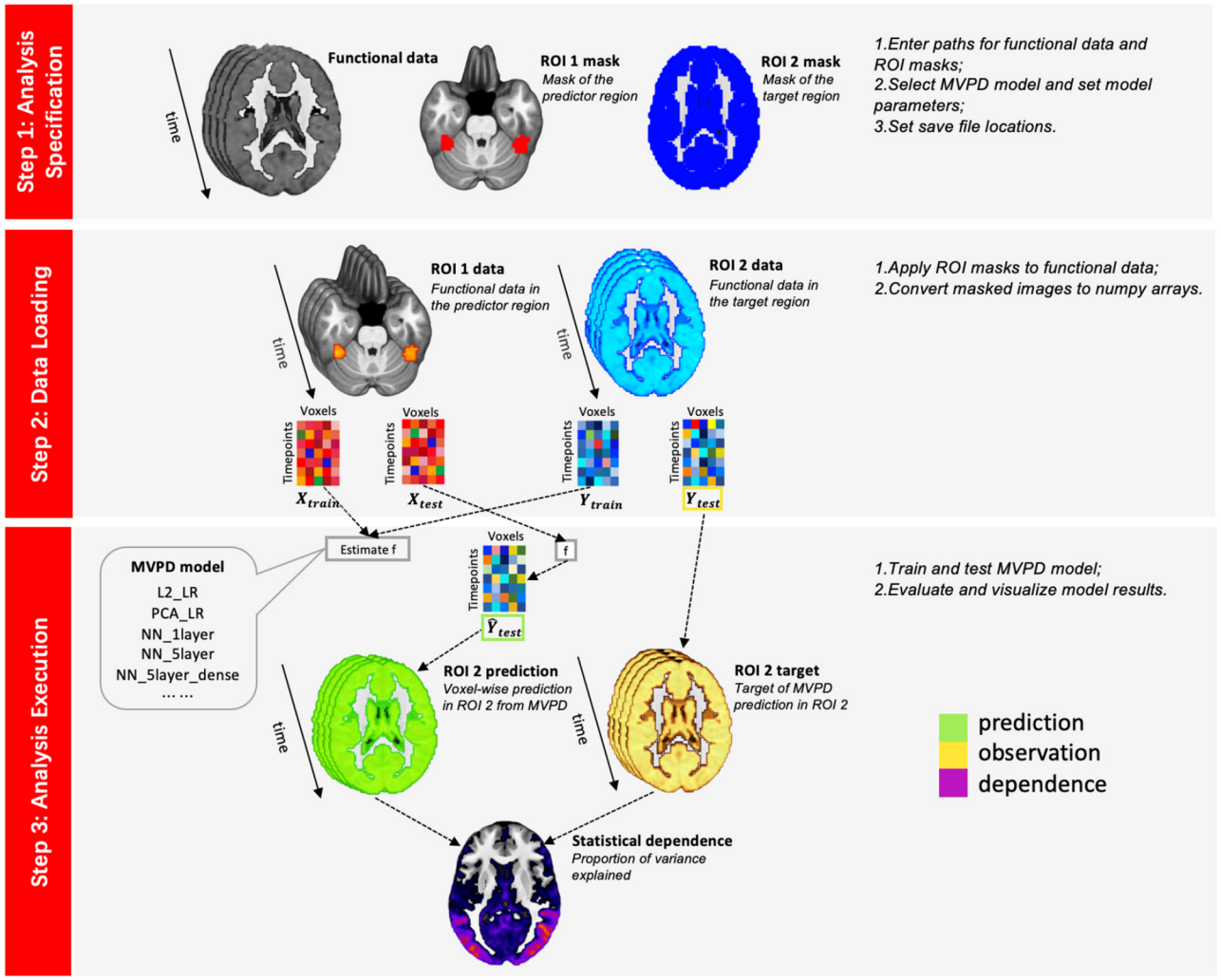


FIGURE 1 | PyMVPD workflow. Analyzing data with the PyMVPD toolbox consists of three steps: (1) analysis specification; (2) data loading; and (3) analysis execution. In step 1, users are required to specify the functional data, masks for both the predictor and the target region, as well as the type of MVPD model to perform the following analyses. These details can be specified by editing the Python script *run_MVPD.py*. Next, users can proceed with step 2, which loads neuroimaging data and converts it to a suitable format. Finally, step 3 runs the MVPD model and generates the analysis results, which are saved in a user-specified directory.

but StudyForrest includes over 2 h of data for each individual participant, making it ideal for MVPD.

The logic of MVPD is as follows. Suppose that we want to calculate the statistical dependence between two brain regions. MVPD will learn a function that, given the response pattern in one region (the "predictor" region), generates a prediction of the response pattern in the other (the "target" region). Let us consider an fMRI scan with m experimental runs. We denote the multivariate timecourses in the predictor region by $X_1, ..., X_m$. Each matrix X_i is of size $n_X \times T_i$, where n_X is the number of voxels in the predictor region, and T_i is the number of timepoints in the experimental run i. Analogously, $Y_1, ..., Y_m$ denote the multivariate timecourses in the target region, where each matrix Y_i is of size $n_Y \times T_i$, and n_Y is the number of voxels in the target region.

As a first step, the data is split into a training subset and a test subset. It is important that the training and test subsets are independent. Since fMRI timeseries are characterized by temporal autocorrelation, it is best to not use timepoints from one run for training and adjacent timepoints from the same run for testing. A common approach is to use leave-one-run-out cross-validation: this is the approach implemented by default in the PyMVPD toolbox. For each choice of an experimental run i, data in the remaining runs is concatenated as the training set

$$D_i = \{(X_1, Y_1), ..., (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), ..., (X_m, Y_m)\},$$

while data $D_i = \{(X_i, Y_i)\}$ in the left-out run i is used as the test set.

For convenience, we will denote with X_{train} and Y_{train} the concatenated training data in the predictor region and in the target region, respectively. The training data is used to learn a function f such that

$$Y_{train} = f(X_{train}) + E_{train},$$

where E_{train} is the error term. In the current implementation, the response pattern in the target region at a given time is predicted from the response pattern in the predictor region at the same time. However, models that integrate the responses in the predictor region across multiple timepoints are a straightforward extension. Once the function f has been estimated, we use it to generate predictions of the responses in the target region $\hat{Y}_{test}$ given the responses in the predictor region during the test run:

$$\hat{Y}_{test} = f(X_{test}).$$

Finally, the accuracy of the prediction is computed. In the PyMVPD toolbox, we provide a measure of predictive accuracy by calculating the voxelwise proportion of variance explained. For each voxel j in the target region, variance explained is calculated as:

$$\text{varExpl}(j) = 1 - \frac{var[Y_{test}(j) - \hat{Y}_{test}(j)]}{var[Y_{test}(j)]},$$

where $\hat{Y}$ are the predicted voxelwise timecourses. The values varExpl(j) are then averaged across voxels in the target region and across cross-validation runs to obtain a single measure varExpl. In addition, the PyMVPD toolbox is designed to allow for customized measures of accuracy (more details will be provided in the following sections).

2.2. The PyMVPD Toolbox

PyMVPD is a Python-based toolbox that implements the MVPD analysis pipeline. This software package is freely available at https://github.com/sccnlab/PyMVPD.git. Artificial neural network models are built using PyTorch—for users who are only interested in linear regression models, or who would like to avoid the complexities of a PyTorch installation, we have also provided a lite version (PyMVPD_LITE) at https://github.com/sccnlab/PyMVPD_LITE.git, for which PyTorch is not required. PyMVPD is based on a simple workflow that consists of three steps: analysis specification, data loading, and analysis execution, with the latter step including the following sub-steps: dimensionality reduction (if requested), model estimation, and model evaluation. Models are trained and tested using k-fold cross validation (where k is a parameter specified by the user).

2.2.1. Preliminaries

Prior to MVPD analysis, the fMRI data at hand should have already undergone standard preprocessing steps, such as registration, normalization and denoising. Denoising is an essential component of preprocessing: measures of statistical dependence are susceptible to noise (Ciric et al., 2017). The preprocessed fMRI data should be in NIfTI file format. Next, the user should create brain masks of the predictor region ("ROI 1") and the target region ("ROI 2"), also in NIfTI file format.

2.2.2. Step 1—Analysis Specification

During analysis specification, the user enters all necessary information to perform the analysis into the script "run_MVPD.py". Information is organized into two variables: "inputinfo", and "params". The variable "inputinfo" contains the paths to the input data as well as the locations to which the results will be saved (the complete list of required values can be found here https://github.com/sccnlab/PyMVPD#required-input-information). The variable "params" contains all details about the analysis, including the type of cross-validation (e.g., leave-k-run-out), the type of dimensionality reduction chosen (if any), the number of dimensions selected, the type of model of statistical dependence, and other hyperparameters of the model (for example, the amount of regularization for regularized regression models, or the neural network architecture for neural network models). We include an overview of the key options available in the Section 2.2.4. A complete description of all parameters would not fit within the limits of this article, therefore it is reported at this page: https://github.com/sccnlab/PyMVPD#list-of-model-parameters (along with the default values for each parameter). Since all parameters for the analysis are specified by the user in step 1, before the analysis is launched, and all results and logs are automatically saved to a user-specified folder, PyMVPD jobs can be launched on computer clusters as batch jobs, without the need to use interactive jobs.

2.2.3. Step 2—Data Loading

The second step of PyMVPD is the loading and processing of input data. Before running the chosen MVPD model, values from the functional data are extracted using masks specified in step 1, and transformed into numpy arrays in preparation for the following analyses. To accomplish this step, the user can execute the line of code *data_loading.load_data(inputinfo)*.

2.2.4. Step 3—Analysis Execution

Once the analysis details have been specified and the data is loaded, the third step executes the analysis, estimating the statistical dependence between brain regions and reporting the accuracy of predictions in independent data. To perform step 3, users can call the function *model_exec.MVPD_exec(inputinfo, params)*, which will estimate the MVPD model, compute the model's performance, and save the results to the folder specified in step 1.

It is important to note that during the implementation of PyMVPD, users only need to interface with the analysis specification script in the first analysis step (e.g., run_MVPD.py). Then, the following two analysis steps will run automatically and the users are not required to interface with any of them. This default setting makes it easier to run the toolbox on computer clusters.

Logging information is saved as a text file named by "TIMESTAMP_log.txt" under the directory where users specify

to save results in the first analysis step. The log file contains information about input data that have been used for the analysis, MVPD model parameters, and the version of the toolbox.

To ensure the toolbox is installed properly and to verify it works, we have included tests for users to run before performing formal analyses. Users can find the test script "run_MVPD_test.py" under the exp/ folder. The test script attempts to replicate on the user's machine the analyses in our manuscript that used FFA as seed using data from subject sub-01, and calculates for each of the five example models the correlation between the variance explained values we obtained and the values obtained with the user's installation across the whole brain. If the correlation values are below 0.95 for any of the model types, the test script returns a warning notifying the user that the results they obtained do not match the benchmarks, specifying which of the models produced results that differed from our reference results. The results of the tests are saved in the folder exp/testresults/, so that the tests can be executed as a batch script on a computer cluster.

Since executing all analyses can take a substantial amount of time, in addition to the "run_MVPD_test.py" script we have included scripts to test individual models. This way, users can test just the type of model they are interested in using. These tests for individual models are also included in the exp/ folder, with the names "run_MVPD_PCA_LR.py", "run_MVPD_L2_LR.py", "run_MVPD_NN_1layer.py", "run_MVPD_NN_5layer.py", and "run_MVPD_NN_5layer_dense.py". In future extensions of the toolbox, we plan to introduce more finer-grained tests for individual functions.

Below, we provide an overview of the available options for the different stages of analysis execution (dimensionality reduction, model estimation, model evaluation). Users can select what options to use for their analysis by editing the file "run_MVPD.py" (as noted in the Section 2.2.4).

2.2.4.1. Dimensionality Reduction

The PyMVPD toolbox offers the option to perform dimensionality reduction on the input data before estimating models of statistical dependence. Dimensionality reduction can be desirable because reducing the dimensionality of the input data leads to a corresponding reduction in the number of parameters of the models, mitigating the risk of overfitting. Two dimensionality reduction approaches are included in the toolbox: principal component analysis (PCA), implemented with sklearn.decomposition.PCA, and independent component analysis (ICA), implemented with sklearn.decomposition.FastICA. In the current implementation of the toolbox, the number of dimensions needs to be entered manually by the user (the default value is 3), but the toolbox is designed to accommodate custom dimensionality reduction functions, offering the possibility to include a nested cross-validation approach for the selection of the number of dimensions (this option may be implemented as a core part of the toolbox in future releases). In particular, for applications in which the choice of the number of dimensions has meaningful theoretical implications, we recommend implementing a custom PCA function that uses Minka's MLE algorithm to select the number of dimensions based on the data. For some model types, dimensionality reduction might not offer additional benefits. In particular, when using artificial neural network models, the neural networks can themselves perform dimensionality reduction as needed—the desired amount of dimensionality reduction can be regulated by choosing the appropriate size of the hidden layer (or layers). Hidden layers with a smaller number of hidden nodes correspond to greater data compression.

2.2.4.2. Model Estimation

2.2.4.2.1. Linear Regression Models Linear regression attempts to model the relationship between a dependent variable and one or more explanatory variables by fitting a linear function to observed data. Specifically, we view the multivariate timecourses in the predictor region X as the explanatory variable and the multivariate timecourses in the target region Y as the dependent variable. The MVPD mapping f can be modeled with multiple linear regression

$$Y_{train} = B_{train}X_{train} + E_{train},$$

where B_{train} is the vector of parameters and E_{train} is the error vector.

A large number of parameters as compared to a relatively small dataset can lead regression models to overfit the data. That is, the model learns a function that corresponds too closely to the particular training set and therefore fails to fit unseen data, resulting in poor predictive accuracy during testing. To mitigate this issue, we provide the option to choose either Lasso or Ridge regularization, setting "params.reg_type" to either "Lasso" or "Ridge". The strength of regularization can be either set manually using the parameter "params.reg_strength" (the default value is 0.001), or automatically thanks to the use of nested cross-validation (Golub et al., 1979). When choosing to set the regularization parameter manually, it is of fundamental importance to decide the value of the parameter a-priori. Performing the analysis with multiple choices of the regularization parameter and selecting the one that yields the best results is a form of circular analysis, and will lead to false positive inflation. To perform automatic selection, we offer the option to use Ridge regularization determining the regularization parameter with a nested cross-validation loop. This can be achieved setting "params.reg_type" to "RidgeCV". By default, the optimal regularization value is chosen among 0.001, 0.01, and 0.1, users can specify a different set of regularization values to test by setting the parameter "params.reg_strength_list". However, it is important to note that automatic selection of the regularization parameter may lead to longer computation times for the analyses.

2.2.4.2.2. Neural Network Models In addition to linear regression models, we introduce an extension of MVPD in which the statistical dependence between brain regions can be modeled using artificial neural networks. In this approach, the multivariate patterns of response in the predictor region are used as the input of a neural network trained to generate the patterns of response in the target region. In PyMVPD, all neural network models are trained using stochastic gradient descent (SGD) on a mean square error (MSE) loss by default. Batch normalization

is applied to the inputs of each layer. Additionally, users should set the following hyperparameters for the chosen neural network: number of hidden units in each layer, number of layers, learning rate, weight decay, momentum, mini-batch size, and number of epochs for training. We provide standard fully-connected feedforward neural network architectures ("NN_standard") and fully-connected feedforward neural network architectures with dense connections ("NN_dense", Huang et al., 2017) for users to choose. Both architectures only consist of linear connections between layers without introducing non-linear activation functions. Users are welcome to build their own neural network models with customized functions.

Note that functional MRI data has temporal dependencies. That is, the amount of response in a voxel in a volume acquired at a given timepoint is not entirely independent of the amount of response in that voxel in the previous timepoint. This non-independence can potentially affect models of the interactions between brain regions, including standard functional connectivity as well as MVPD. In order to mitigate the effect of non-independence, the neural network models in PyMVPD adopt a strategy borrowed from deep Q-learning. In deep Q-learning there is a similar problem: the actions taken by a reinforcement learning agent, the resulting states of the environment, and the rewards are non-independent across adjacent timepoints. To mitigate this problem, actions and the resulting states are logged to a "replay memory"; the policy network is then trained on a batch sampled randomly from the replay memory, so that the actions, states and rewards in each batch are more independent (see Fan et al., 2020). PyMVPD neural network models use the same strategy: each training batch contains datapoints collected at a randomly sampled set of timepoints. As a consequence, the set of datapoints within a batch are more independent than if they had been sampled consecutively.

2.2.4.2.3. Searchlight Analysis Previous MVPD studies included searchlight-based analyses (Anzellotti et al., 2017a). The results of searchlight analyses can be contingent on the use of a sphere as the searchlight shape, and on the choice of a particular radius. To avoid this, we recommend using multi-output models instead: users interested in mapping the statistical dependence between one region and the rest of the brain can use a whole-brain mask as the target region (as we have done in the present work).

2.2.4.3. Model Evaluation

To measure the predictive accuracy of the MVPD model after execution, we included code to calculate variance explained following two different approaches. In one approach, the variance explained values are left unthresholded, and thus can range between $-\infty$ and 1. This can be helpful to identify cases in which there is a clear mismatch between the target and the prediction. However, since negative values of variance explained are difficult to interpret in terms of their neuroscientific implications, and since very negative outliers in individual participants can conceal voxels with positive variance explained in most participants, we additionally implemented a function to set negative variance explained to zero, indicating that the model failed to predict the responses in a given voxel for a given participant.

Notably, even when setting negative values of the variance explained to zero, the variance explained approach is more stringent than computing Pearson correlation between the model predictions and the observations. For example, in the presence of predictions that match the observations in terms of their patterns, but show a large difference in the means, Pearson correlation would be very high, while variance explained would be zero.

Statistical significance can be computed by performing a permutation test on the unthresholded variance explained values. Alternatively, phase resampling of the responses in the target of prediction could be used to construct the null distribution (see Liu and Molenaar, 2016). In addition, comparisons between the predictive accuracy of different predictor regions or different types of models can be done using non-parametric statistical tests on the differences between their proportions of variance explained (thresholded or unthresholded). This was the approach we adopted in the experimental application of PyMVPD in this article.

We assessed the statistical significance across participants with statistical non-parametric mapping (Nichols and Holmes, 2002) using the SnPM13 software, using FWE-correction at the voxel level to control for multiple comparisons (http://warwick.ac.uk/snpm). More specifically, to identify significant differences between two models, we first computed the average variance explained across cross-validation iterations for each voxel and for each model, and then we computed differences between these averages for the two models, obtaining one difference map for each participant. Finally, these difference maps were entered in SnPM13 following the steps described in this tutorial: https://warwick.ac.uk/fac/sci/statistics/staff/academic-research/nichols/software/snpm/man/exnew, and significance was computed selecting the option "MultiSub: One Sample *t*-test on diffs/contrasts". When computing statistical significance, it is important to consider the spatial autocorrelation of fMRI data: the measurements from nearby voxels tend to be correlated. Some types of correction for multiple comparisons (e.g., cluster correction) can be susceptible to spatial autocorrelation, when using such methods, underestimating the spatial autocorrelation may lead to excessively liberal statistical thresholds. When there is not complete confidence that spatial autocorrelation can be correctly estimated, we recommend using thresholds corrected at the voxel level.

Importantly, if negative values of variance explained are set to zero, the use of standard statistical tests (such as *t*-tests) to establish significance can lead to exceedingly liberal thresholds. Recent work has investigated in depth this problem in the context of classification accuracy (Allefeld et al., 2016; Hirose, 2021), introducing new statistics that can be used to address this issue. Future work may lead to the development of approaches to implement FWE correction for these statistics, making it possible to apply them to whole-brain analyses controlling for multiple comparisons. In the meantime, we recommend either using raw variance explained values (without setting negative values to zero), or performing statistical tests on subtractions between the variance

explained values obtained with different model types or different brain regions.

2.3. Application to Experimental fMRI Data

2.3.1. Data Acquisition and Preprocessing

As a demonstration of the use of PyMVPD, we analyzed fMRI data of 15 participants (age range 21–39 years, mean 29.4 years, 6 females) watching a movie, from the publicly available *StudyForrest* dataset (http://studyforrest.org). Functional data were collected on a whole-body 3 Tesla Philips Achieva dStream MRI scanner equipped with a 32 channel head coil. The BOLD fMRI responses at the resolution of 3×3×3 mm were acquired using a T2*-weighted echo-planar imaging sequence. Complete details can be found in Hanke et al. (2016).

The dataset includes a movie stimulus session, collected while participants watched the 2-h audio-visual movie "Forrest Gump". The movie was cut into eight segments, and each segment lasted approximately 15 min. All eight segments were presented to participants in chronological order in eight separate functional runs. Additionally, the dataset includes an independent functional localizer that can be used to identify category-selective regions (Sengupta et al., 2016). During the category localizer session, participants viewed 24 unique gray-scale images from each of six stimulus categories: human faces, human bodies without heads, small artifacts, houses, outdoor scenes, and phase scrambled images. Each participant was presented with four block-design runs and a one-back matching task.

All fMRI data was preprocessed using fMRIPrep (https://fmriprep.readthedocs.io/en/latest/index.html). Anatomical images were skull-stripped with ANTs (http://stnava.github.io/ANTs/), and segmented into gray matter, white matter, and cerebrospinal fluid using FSL FAST. Functional images were corrected for head movement with FSL MCFLIRT (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MCFLIRT), and were subsequently coregistered to their anatomical scan with FSL FLIRT. Data of one participant was excluded because it could not pass the fMRIPrep processing pipeline. For the remaining 14 participants, we removed noise from the data with CompCor (Behzadi et al., 2007) using 5 principal components extracted from the union of cerebrospinal fluid and white matter. Regions of no interest for the cerebrospinal fluid and white matter were defined individually for each participant.

2.3.2. ROI Definition

In each individual participant, seed regions of interest (ROIs) in the fusiform face areas (FFA) as well as the parahippocampal place areas (PPA) were defined using the first block-design run from the functional localizer. We performed whole-brain first level analyses on each participant's functional data by applying a standard general linear model with FSL FEAT (Woolrich et al., 2001). Next, we identified the peak voxels with the highest *t*-values for the contrast between the preferred category and other categories (i.e., FFA contrast: faces > bodies, artifacts, scenes, and scrambled images; PPA contrast: scenes > faces, bodies, artifacts, and scrambled images). We generated spheres of 9 mm radius centered in the peaks. Finally, the voxels within spheres from the left and right hemispheres were combined, and the 80 voxels with the highest *t*-values were selected (this is a common choice in neuroimaging studies, see Skerry and Saxe, 2014; Kliemann et al., 2018).

We additionally created a group-average gray matter mask using the gray matter probability maps generated during preprocessing, with a total of 53,539 voxels, that was used as the target of prediction.

2.3.3. PyMVPD Analysis

Using the PyMVPD toolbox, we estimated the multivariate pattern dependence between each ROI (FFA/PPA) and the gray matter using five example MVPD models: **L2_LR**, **PCA_LR**, **NN_1layer**, **NN_5layer**, and **NN_5layer_dense**. **L2_LR** is a linear regression model with Ridge (L2) regularization. The regularization strength was set to be 0.001. **PCA_LR** is a linear regression model that applies dimensionality reduction on input data with PCA using three principal components. **NN_1layer** and **NN_5layer** are fully-connected feedforward neural networks derived from the "NN_standard" architecture with one hidden layer and five hidden layers, respectively. Under the "NN_dense" architecture, **NN_5layer_dense** is a fully-connected feedforward neural network with dense connections and five hidden layers. For all the neural network models, we set the number of hidden units in each hidden layer to be 100. Each network was trained with a batch size of 32, a learning rate of 0.001, and a momentum of 0.9 with no weight decay.

For both FFA and PPA, we took the 80 voxels in the seed ROI as the predictor region, and the 53,539 voxels in the gray matter as the target region. For each MVPD model, 7 of the 8 movie runs were used for training, and the remaining run was used for testing. This leave-one-run-out procedure was repeated 8 times by leaving aside each possible choice of the left-out run. We then calculated the variance explained for each voxel in the target region with all five MVPD models in the left-out data.

The proportion of variance explained for each seed region and model was computed for each voxel in gray matter, negative values of variance explained were set to 0. Next, we compared the overall predictive accuracy in each pair of MVPD models. For each participant, the proportion of variance explained by each model was averaged across all voxels in the gray matter, and across all cross-validation folds. The difference between the average variance explained by the two models was computed for each participant, and the significance was assessed with a one-tailed *t*-test across participants—*p*-values were Bonferroni corrected for all 20 comparisons (since one-tailed tests were used, comparisons in both directions were counted in the correction).

In addition to testing the models' overall predictive accuracy, we sought to compare their accuracy at the level of individual voxels. First, we performed a voxelwise comparison of neural network models vs. linear regression models. To do this, for each voxel, we calculated the average variance explained across neural network (NN) models, and we subtracted the average variance explained across linear regression (LR) models. We computed statistical significance across participants with statistical non-parametric mapping using the SnPM13 software, obtaining pseudo-t statistics for each voxel. Then, we identified

voxels where neural network models significantly outperformed linear regression models, at a familywise error (FWE) corrected threshold of $p < 0.05$ (voxelwise FWE-correction was used). Next, we performed finer-grained analyses, focusing on the better performing NN models. In particular, we tested all pairwise comparisons between individual NN models. As in the previous analysis, significance was computed using SnPM13, using a voxelwise FWE-corrected threshold of $p < 0.05$. We used Bonferroni correction to control for the number of multiple comparisons.

Even in regions where different models are not significantly different, qualitative differences might reveal large-scale patterns that could help users in the selection of a particular model. Given this consideration, we aimed to provide a qualitative evaluation of the relative performance of different models across the brain. To this end, for each voxel in the gray matter, we first selected the model yielding the highest proportion of variance explained in that voxel (averaged across participants) and specified that model as the best model for that voxel. Then, we obtained a conservative measure of the extent to which the model outperformed the other models by calculating the lowest *t*-value among all comparisons between the best model and all other models. As these results are qualitative in nature, they are shown in the **Supplementary Material** (**Supplementary Figure 6**).

3. RESULTS

In line with previous studies using MVPD (Anzellotti et al., 2017a), our implementation of PyMVPD identified the expected patterns of statistical dependence between FFA and PPA and other brain regions (see **Figure 2A** for a visualization of the seed regions in one representative participant). Across multiple model types, when using FFA as seed, regions showing high variance explained included other face-selective regions, and when using PPA as seed, regions showing high variance explained included other scene selective regions (**Supplementary Figures 1–5**, peak coordinates for face-selective regions and scene-selective regions were determined with Neurosynth, https://www.neurosynth.org/). In subsequent analyses, we focused on comparing the performance of different models, first in terms of their overall accuracy (averaged across the entire brain), and then at the level of individual voxels.

3.1. Comparing the Average Performance of Different Models

To compare the overall predictive accuracy across different MVPD models, the proportion of variance explained for each model was averaged across the whole brain. Then, we performed pairwise comparisons among all five example models. For each pair of models, we subtracted the variance explained varExpl of one model from that of another one. This procedure yielded a difference value for each participant, and we conducted a one-sample one-tailed *t*-test on the difference values across all 14 participants using SnPM. All *p*-values were Bonferroni corrected for 20 multiple comparisons. Results are shown in **Figure 2B** as difference matrices for FFA and PPA, respectively.

Overall, models based on artificial neural networks outperformed standard linear regression models. Linear regression based on principal component analysis (**PCA_LR**) showed the worst predictive accuracy while **NN_5layer_dense** proved to be the best predicting model. More precisely, using FFA as seed region, **L2_LR** showed a significantly higher average variance explained than **PCA_LR** [$t_{(13)} = 6.05$, $p = 0.0004$ corrected]. Both **NN_1layer** and **NN_5layer** significantly outperformed **PCA_LR** in terms of average variance explained [NN_1layer: $t_{(13)} = 6.27$, $p = 0.00028$ corrected; NN_5layer: $t_{(13)} = 6.46$, $p = 0.00022$ corrected]. **NN_5layer_dense** revealed a significantly higher average variance explained than **L2_LR** [$t_{(13)} = 7.72$, $p < 0.0002$ corrected] and **PCA_LR** [$t_{(13)} = 6.37$, $p = 0.00024$ corrected]. Using PPA as seed region, all the other models showed significantly better predictive performance than **PCA_LR** [L2_LR: $t_{(13)} = 6.55$, $p < 0.0002$ corrected; NN_1layer: $t_{(13)} = 5.79$, $p = 0.00062$ corrected; NN_5layer: $t_{(13)} = 5.49$, $p = 0.00104$ corrected; NN_5layer_dense: $t_{(13)} = 6.68$, $p < 0.0002$ corrected]. In addition, **NN_5layer_dense** also significantly outperformed **L2_LR**, **NN_1layer** and **NN_5layer** [L2_LR: $t_{(13)} = 6.30$, $p = 0.00028$ corrected; NN_1layer: $t_{(13)} = 3.45$, $p = 0.04308$ corrected; NN_5layer: $t_{(13)} = 3.73$, $p = 0.02522$ corrected]. The rest of the pairwise comparisons did not show significant differences across participants ($p > 0.05$ corrected).

3.2. Comparing the Performance of Different Models at the Level of Individual Voxels

To further understand the relative accuracy of different models in different brain regions, we tested the relative performance of neural network models (NN) to the performance of linear regression (LR) models. In particular, we averaged the variance explained for each voxel across the three NN models (**NN_1layer**, **NN_5layer**, **NN_5layer_dense**), and the two LR models (**L2_LR**, **PCA_LR**), respectively. Given the higher predictive accuracy of NN models over LR models when averaged across the whole brain (**Figure 2B**), we also expected NN models to outperform LR models in several brain regions. We tested this hypothesis by calculating the difference in predictive accuracy between average NN and LR models for each voxel in each participant, and then computed the statistical significance across participants. As expected, the resulting SnPM t-map shown in **Figure 3** revealed a large portion of the gray matter that was better predicted by the average NN models rather than the average LR models using FFA or PPA as seed region. NN models did not achieve higher predictive accuracy in the seed regions—this is to be expected, since a very simple model such as the identity function would be sufficient in these regions. By contrast, responses in the other category-selective regions (i.e., face-selective regions: OFA, STS, ATL; scene-selective regions: RSC, TOS) were better predicted by the average NN models over the average LR models when using the seed region of the matching category (**Figure 3**).

Next, we investigated in more detail the relative voxel-wise predictive accuracy among the three NN models. To do

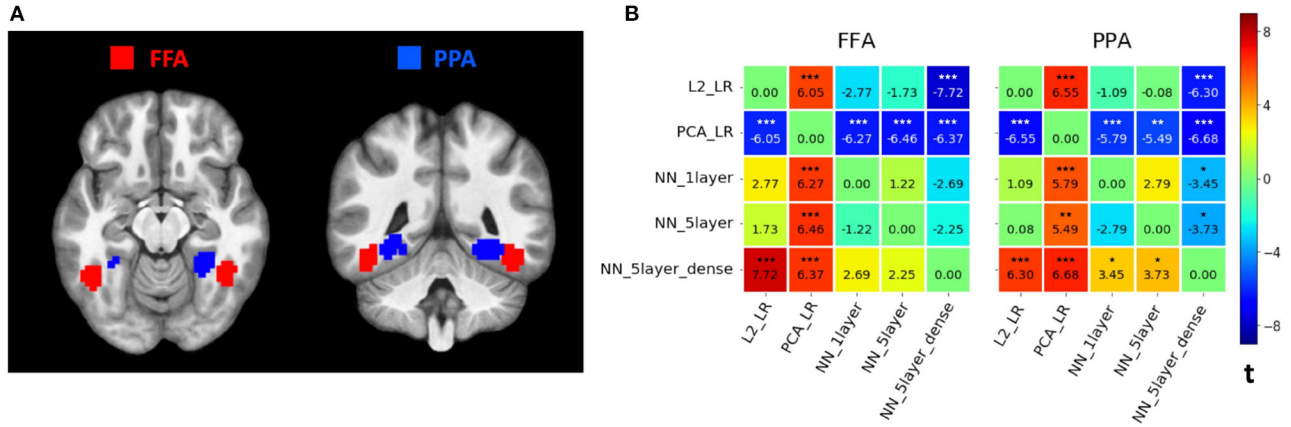


FIGURE 2 | (A) Seed regions. Example seed regions for one representative participant. The fusiform face areas (FFA) are shown in red, and the parahippocampal place areas (PPA) are shown in blue. Individual ROIs per participant were defined based on first-level t-maps, identifying 9 mm spheres centered on the peaks for the preferred category, and selecting the top 80 voxels with highest *t*-values within the spheres. **(B)** Comparison between different MVPD models. For each seed predictor region (left: FFA; right: PPA), we plotted the difference matrix of *t*-values across the five example MVPD models. As a measure of overall predictive accuracy, the average proportion of variance explained varExpl was computed across the whole brain for each model per participant. For each pair of models, we subtracted the varExpl of one model from another one, obtaining the pairwise difference values. Finally, we conducted a one-tailed *t*-test on the difference values across all 14 participants. The corresponding *t*-values were entered into the difference matrix, indicating the extent to which one model outperformed another one in terms of the overall predictive accuracy using the seed region as the predictor. Stars above *t*-values indicate the statistical significance (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$; Bonferroni-corrected).

this, we calculated the difference values of variance explained between each pair of NN models (6 pairs in total). Statistical significance was computed using SnPM and all *p*-values were Bonferroni corrected for 6 multiple comparisons. Due to controlling for multiple comparisons both across voxels (with a FWE-corrected voxelwise threshold determined with SnPM) and across multiple model comparisons (thus further dividing the threshold by 6), this analysis is very stringent. Nonetheless, the analysis did reveal some loci of significant differences between the models (**Figure 4**). Using FFA as seed region, the insula was significantly better predicted by **NN_5layer** over **NN_5layer_dense** (**Figure 4A**), and a region in left parietal cortex was significantly better predicted by **NN_5layer** over **NN_1layer** (**Figure 4B**). Using PPA as seed region, a region in the cerebellum showed significant higher predictive accuracy by **NN_1layer** than **NN_5layer**, by **NN_5layer_dense** than **NN_1layer**, and by **NN_5layer_dense** than **NN_5layer**. Additional, smaller loci showing significant differences are reported in **Supplementary Table 1** (FFA) and **Supplementary Table 2** (PPA).

Finally, since qualitative differences that do not pass significance might still be helpful for users interested in choosing a model, we generated a map that visualizes the best performing model for each voxel, and the extent to which the best model outperforms the other models (**Supplementary Figure 6**). Specifically, we assigned different colors to each model (**L2_LR**: green; **PCA_LR**: blue; **NN_1layer**: red; **NN_5layer**: yellow; **NN_5layer_dense**: purple). The color of each voxel was set to the color of the model that performed best at predicting that voxel's responses, and the color's saturation was set proportionally to the lowest *t*-value from all pairwise comparisons between models. In other words, more saturated colors appear in voxels for which the difference between the best model and the runner-up model is greater. Together, the voxelwise analyses revealed that there isn't a single best model for all voxels, instead, different voxels are best predicted by different models.

4. DISCUSSION

In this article, we have introduced PyMVPD, a Python-based toolbox for multivariate pattern dependence (MVPD). MVPD is a novel technique that investigates the statistical relationship between the responses in different brain regions in terms of their multivariate patterns of response (Anzellotti et al., 2017a). Previous studies have shown that this approach brings higher sensitivity in detecting statistical dependence than standard functional connectivity (Anzellotti et al., 2017a,b). However, given the complex nature of the analysis, the implementation of MVPD can be an obstacle to its wider application. PyMVPD enables researchers to perform complex MVPD analyses with a few lines of easily readable Python code, therefore, it makes MVPD more accessible to a broader community of researchers.

PyMVPD provides users with a flexible analysis framework to study the multivariate statistical dependence between brain regions. Users can choose whether or not to use dimensionality reduction, and if dimensionality reduction is selected, PyMVPD offers a choice between principal component analysis (PCA) and independent component analysis (ICA). Furthermore, PyMVPD permits the use of a variety of models to study the multivariate statistical dependence between brain regions. In addition to

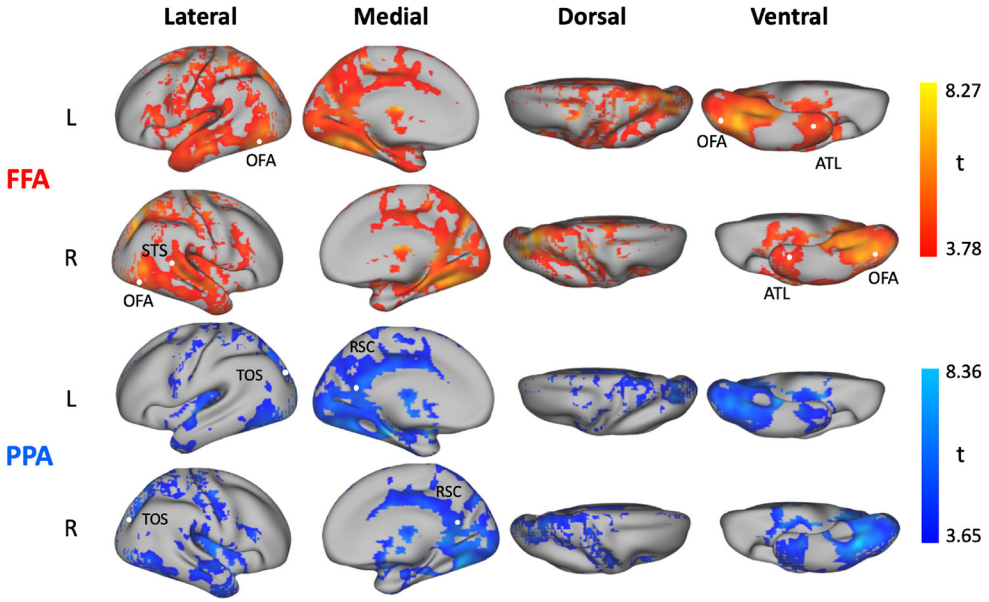


FIGURE 3 | Comparison between neural network (NN) models and linear regression (LR) models. Statistical t-maps computed across subjects from the voxelwise difference between the average variance explained predicted by three NN models (**NN_1layer, NN_5layer, NN_5layer_dense**), and the average variance explained predicted by two LR models (**L2_LR, PCA_LR**) with FFA (top) and PPA (bottom) as predictor ROIs, respectively. The SnPM threshold corrected at $p < 0.05$ FWE is 3.78 using FFA as predictor and is 3.65 using PPA as predictor.

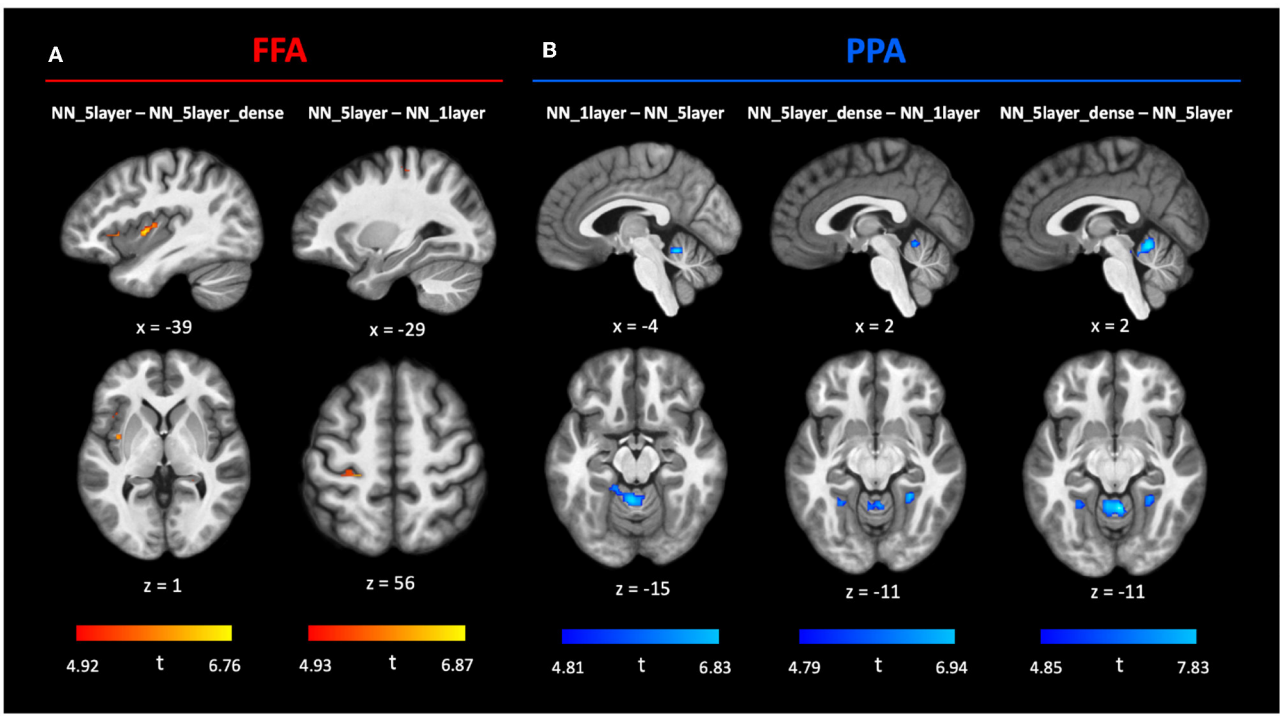


FIGURE 4 | Comparison between MVPD neural network (NN) models. Statistical t-maps computed across subjects from the pairwise difference between the variance explained predicted by three neural network models (**NN_1layer, NN_5layer, NN_5layer_dense**) with FFA **(A)** and PPA **(B)** as predictor ROIs, respectively. The SnPM p-values were Bonferroni corrected for all 6 comparisons. We showed interesting brain regions that were better predicted by one NN model than the other NN model at $p < 0.05$ FWE after Bonferroni correction. The full NN model comparison results can be found in **Supplementary Table 1** (FFA) and **Supplementary Table 2** (PPA).

the standard linear regression models that have proven to be effective in the previous literature (Anzellotti et al., 2017a), we make artificial neural networks available for connectivity research through an integration with PyTorch. The artificial neural network version of MVPD implemented in PyTorch is introduced for the first time in this article. We demonstrate that the neural network implementations of MVPD outperform the previously published version based on PCA in most brain regions. Example code is provided for three neural network architectures. In addition, users can choose other architectures with different numbers of hidden units and of layers by changing the parameter settings. For applications that require models beyond the family of options already available in PyMVPD, the toolbox is designed so that it is straightforward to program custom architectures and to integrate them with the other scripts.

In the experimental applications described in this work, we tested PyMVPD using the StudyForrest dataset, with the FFA and PPA as seed regions. The results revealed interactions between these seed regions and the rest of the brain during movie-watching, following a pattern that is consistent with the previous literature. Category-selective peaks identified with Neurosynth fell within the MVPD maps for the corresponding category. Overall, artificial neural networks outperformed linear regression models in terms of the predictive accuracy for statistical dependence. Importantly, this is not a trivial consequence of the fact that the artificial neural networks are more complex. In fact, MVPD trains and tests models with independent subsets of the data, and models with more parameters do not necessarily perform better at out-of-sample generalization.

Interestingly, no single model outperformed all others in every voxel. In particular, the **NN_5layer** outperformed other models at predicting responses in the insula and parietal regions using the FFA seed as predictor. By contrast, **NN_5layer_dense** outperformed other models at predicting cerebellar responses given PPA inputs. A qualitative analysis revealed large, contiguous cortical regions in which one model type outperformed the others (**Supplementary Figure 6**). Taken together, these results indicate that the statistical dependence between different sets of regions might be best characterized by different models. Why would this be the case? It is expected that the interactions between different sets of brain regions implement different kinds of computations. For example, the computations implemented by the interaction between the fusiform face area (FFA) and the occipital face area (OFA)—hypothesized to be upstream of FFA in a hierarchy of visual processing—are likely to be different from the computations implemented by the interaction between the FFA and frontal cortex regions involved in attention. We hypothesize that such differences in the underlying computations could lead to differences in terms of which neural network architectures yield the best models of between-region interactions.

The present results have broader implications for the study of statistical dependence between brain regions: in the literature on brain connectivity, the focus has been largely placed on whether or not two brain regions interact. However, a key direction for future research consists in investigating not only whether two regions interact, but also how they interact. The observation that the statistical dependence between the seed regions and different voxels were best captured by different models suggests that PyMVPD could be used to make progress in this direction.

To pursue goals such as this, PyMVPD is designed to be easily customized and extended. In addition to the five example models (i.e., **L2_LR**, **PCA_LR**, **NN_1layer**, **NN_5layer**, **NN_5layer_dense**) implemented in this article, PyMVPD allows users to build their own MVPD models with customized function components as well as evaluation metrics, making this toolbox an ideal environment to compare the predictive accuracy of different types of models to study the interactions between brain regions.

Installing the full version of PyMVPD requires a working installation of PyTorch, installed compatibly with the version of the CUDA drivers of the GPUs. For users who prefer to avoid this step and do not need to use the neural networks, we make available the LITE version of PyMVPD, that includes only the linear regression models, and does not require PyTorch. The LITE version can be also installed using the Python Package Index (with "pip").

The toolbox offers a variety of different models that can be used to characterize the interactions between brain regions. The selection of a model among the available options can be based on multiple considerations. First, in this study, we found that artificial neural network models were more accurate than the PCA-based linear regression and the L2 linear regression overall. For this reason, when analyzing a comparable amount of data, and when maximum accuracy is needed, we recommend using artificial neural networks. However, models using artificial neural networks require a working Pytorch installation, and the additional accuracy they offer might not be needed for some use cases. In addition, it is essential to note that there is a trade-off between model complexity and model fit: more complex models may not perform well when the amount of data is limited. For this reason, when a smaller number of volumes is available for training, we recommend using the L2 linear regression (Ridge Regression) model, as it offers the additional flexibility of setting the regularization parameter appropriately for the amount of data available. We also note that the optimal model choice may depend not only on the amount of available data, but also on the amount of noise in the data. For this reason, in cases where maximizing the accuracy is essential, we recommend using data from a small subset of participants to test and compare multiple different model choices. The best performing model can then be used to analyze data from the left-out participants. To avoid circularity in the analyses, it is essential to ensure that the data used to select the optimal model are not later reused to estimate the variance explained by that model.

Together with both versions of PyMVPD, we provide step-by-step tutorials on how to calculate MVPD using the toolbox (https://github.com/sccnlab/PyMVPD/blob/main/exp/PyMVPD_Tutorial.ipynb, https://github.com/sccnlab/PyMVPD_LITE/blob/main/exp/PyMVPD_LITE_Tutorial.ipynb). The tutorials are written with Jupyter Notebook, and include sample data as well as the option to plot one's results side by side with the results we computed. This will make it easier for users to check that the toolbox was installed correctly and to confirm that the results match with those we obtained.

Despite the several options available in PyMVPD, the toolbox still has several limitations. For example, functions to automatically select the optimal number of dimensions from the data when using dimensionality reduction have not yet been implemented. In addition, while PyMVPD offers a variety of neural network architectures, including standard feedforward neural networks and DenseNets, other architectures (such as ResNets) are not available, and would require users to develop their own custom code, which can be integrated with the rest of the toolbox. Importantly, we note that the scope of the toolbox is restricted to multivariate analyses of statistical dependence based on MVPD, and as such it does not include other multivariate measure of statistical dependence, univariate measures of statistical dependence such as functional connectivity, nor other multivariate analyses such as decoding or representational similarity analysis. For such analyses, there are several other existing toolboxes that can be used. In particular, users interested in univariate analyses of connectivity may use the Conn toolbox (Whitfield-Gabrieli and Nieto-Castanon, 2012) or GraphVar (Kruschwitz et al., 2015), and users interested in multivoxel pattern analysis (MVPA), including multivariate decoding and representational similarity analysis, may use the PyMVPA toolbox (Hanke et al., 2009), the CoSMoMVPA toolbox (Oosterhof et al., 2016), or "the decoding toolbox" (tdt, Hebart et al., 2015).

A common criticism of methods based on artificial neural networks is that they operate as a black box: it can be difficult to interpret how the neural networks work in terms of cognitively relevant dimensions. Fortunately, an increasing number of techniques are being developed to improve the interpretability of artificial neural networks (Zhang and Zhu, 2018; Li et al., 2021). While additional work will be needed to integrate these techniques with MVPD, the current MVPD framework based on artificial neural networks already offers the benefit of more sensitive detection of statistical dependence as compared to regularized regression, and the opportunity to compare the performance of different model architectures.

The present study focused on the FFA and PPA as seed regions because they have been studied in depth in previous literature. Future studies can extend our results, investigating the application of PyMVPD to other seed regions. The current implementation of PyMVPD is based on simultaneous prediction: responses in the target region at a given time are predicted from responses in the predictor region at the same time. However, other researchers could take advantage of the customization options to use the responses in multiple timepoints in the predictor region to predict the responses in the target region at each timepoint. Finally, the models of statistical dependence implemented by PyMVPD are deterministic. Multivariate probabilistic models that capture the distribution of uncertainty in predictions are in principle possible, but would require large amounts of data for training.

Although PyMVPD was specifically developed for fMRI analysis, the generic design of the framework makes it widely applicable to other data acquisition modalities (i.e., EEG, MEG) across a variety of domains of brain imaging research. We hope that this toolbox removes some of the barriers to the adoption of MVPD, and facilitates the diffusion of multivariate analyses of the interactions between brain regions.

DATA AVAILABILITY STATEMENT

The fMRI data used in this study can be obtained from https://www.studyforrest.org. The PyMVPD toolbox code is available at https://github.com/sccnlab/PyMVPD. Further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Otto-von-Guericke University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MF, CP, and SA: study conception, design, analysis, and interpretation of results. MF and SA: toolbox development and draft manuscript preparation. All authors reviewed the results and approved the final version of the manuscript.

FUNDING

This work was supported by a Startup Grant from Boston College and by NSF Grant 1943862 to SA.

ACKNOWLEDGMENTS

We would like to thank Aidas Aglinskas, Emily Schwartz, and Tony Chen for their comments on a previous version of the manuscript, as well as two reviewers for their constructive feedback. We thank the researchers who contributed to the *StudyForrest* project (Hanke et al., 2016; Sengupta et al., 2016) for sharing their data, and the developers of fmriprep (Esteban et al., 2019) for their assistance with the fmriprep preprocessing pipeline. We would also like to thank Kyle Kurkela for his suggestions about improvements to the code.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fninf.2022.835772/full#supplementary-material

REFERENCES

Allefeld, C., Görgen, K., and Haynes, J.-D. (2016). Valid population inference for information-based imaging: from the second-level t-test to prevalence inference. *Neuroimage* 141, 378–392. doi: 10.1016/j.neuroimage.2016.07.040

Anzellotti, S., Caramazza, A., and Saxe, R. (2017a). Multivariate pattern dependence. *PLoS Comput. Biol.* 13, e1005799. doi: 10.1371/journal.pcbi.1005799

Anzellotti, S., and Coutanche, M. N. (2018). Beyond functional connectivity: investigating networks of multivariate representations. *Trends Cogn. Sci.* 22, 258–269. doi: 10.1016/j.tics.2017.12.002

Anzellotti, S., Fedorenko, E., Kell, A. J., Caramazza, A., and Saxe, R. (2017b). Measuring and modeling nonlinear interactions between brain regions with fMRI. *bioRxiv* 074856. doi: 10.1101/074856

Basti, A., Nili, H., Hauk, O., Marzetti, L., and Henson, R. (2020). Multivariate connectivity: a conceptual and mathematical review. *Neuroimage* 221, 117179. doi: 10.1016/j.neuroimage.2020.117179

Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007). A component based noise correction method (compcor) for bold and perfusion based fMRI. *Neuroimage* 37, 90–101. doi: 10.1016/j.neuroimage.2007.04.042

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., et al. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *Neuroimage* 154, 174–187. doi: 10.1016/j.neuroimage.2017.03.020

Coutanche, M. N., and Thompson-Schill, S. L. (2013). Informational connectivity: identifying synchronized discriminability of multi-voxel patterns across the brain. *Front. Hum. Neurosci.* 7, 15. doi: 10.3389/fnhum.2013.00015

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isil, A. I., Erramuzpe, A., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods.* 16, 111–116. doi: 10.1038/s41592-018-0235-4

Fan, J., Wang, Z., Xie, Y., and Yang, Z. (2020). "A theoretical analysis of deep Q-learning," in *Learning for Dynamics and Control* (PMLR), 486–489.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* 2, 56–78. doi: 10.1002/hbm.460020107

Friston, K. J., Harrison, L., and Penny, W. (2003). Dynamic causal modelling. *Neuroimage* 19, 1273–1302. doi: 10.1016/S1053-8119(03)00202-7

Geerligs, L., Can, C., and Henson, R. N. (2016). Functional connectivity and structural covariance between regions of interest can be measured more accurately using multivariate distance correlation. *Neuroimage* 135, 16–31. doi: 10.1016/j.neuroimage.2016.04.047

Goebel, R., Roebroeck, A., Kim, D.-S., and Formisano, E. (2003). Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and granger causality mapping. *Magnet. Reson. Imaging* 21, 1251–1261. doi: 10.1016/j.mri.2003.08.026

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* 21, 215–223. doi: 10.1080/00401706.1979.10489751

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometr. J. Econometr. Soc.* 37, 424–438. doi: 10.2307/1912791

Greicius, M. D., Krasnow, B., Reiss, A. L., and Menon, V. (2003). Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. U.S.A.* 100, 253–258. doi: 10.1073/pnas.0135058100

Hanke, M., Adelhöfer, N., Kottke, D., Iacovella, V., Sengupta, A., Kaule, F. R., et al. (2016). A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation. *Sci. Data* 3, 160092. doi: 10.1038/sdata.2016.92

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., and Pollmann, S. (2009). PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics* 7, 37–53. doi: 10.1007/s12021-008-9041-y

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Haynes, J.-D., and Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7, 523–534. doi: 10.1038/nrn1931

Hebart, M. N., Görgen, K., and Haynes, J.-D. (2015). The decoding toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Front. Neuroinform.* 8, 88. doi: 10.3389/fninf.2014.00088

Hirose, S. (2021). Valid and powerful second-level group statistics for decoding accuracy: information prevalence inference using the i-th order statistic (i-test). *Neuroimage* 242, 118456. doi: 10.1016/j.neuroimage.2021.118456

Horwitz, B., Grady, C. L., Haxby, J. V., Schapiro, M. B., Rapoport, S. I., Ungerleider, L. G., et al. (1992). Functional associations among human posterior extrastriate brain regions during object and spatial vision. *J. Cogn. Neurosci.* 4, 311–322. doi: 10.1162/jocn.1992.4.4.311

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), 4700–4708. doi: 10.1109/CVPR.2017.243

Kliemann, D., Richardson, H., Anzellotti, S., Ayyash, D., Haskins, A. J., Gabrieli, J. D., et al. (2018). Cortical responses to dynamic emotional facial expressions generalize across stimuli, and are sensitive to task-relevance, in adults with and without autism. *Cortex* 103, 24–43. doi: 10.1016/j.cortex.2018.02.006

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2, 4. doi: 10.3389/neuro.06.004.2008

Kruschwitz, J., List, D., Waller, L., Rubinov, M., and Walter, H. (2015). Graphvar: a user-friendly toolbox for comprehensive graph analyses of functional brain connectivity. *J. Neurosci. Methods* 245, 107–115. doi: 10.1016/j.jneumeth.2015.02.021

Lachenbruch, P. A., and Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics* 10, 1–11. doi: 10.1080/00401706.1968.10490530

Li, X., Xiong, H., Li, X., Wu, X., Zhang, X., Liu, J., et al. (2021). Interpretable deep learning: interpretation, interpretability, trustworthiness, and beyond. *arXiv preprint arXiv:2103.10689*. doi: 10.48550/arXiv.2103.10689

Li, Y., Saxe, R., and Anzellotti, S. (2019). Intersubject MVPD: empirical comparison of fMRI denoising methods for connectivity analysis. *PLoS ONE* 14, e0222914. doi: 10.1371/journal.pone.0222914

Liu, S., and Molenaar, P. (2016). Testing for granger causality in the frequency domain: a phase resampling method. *Multivar. Behav. Res.* 51, 53–66. doi: 10.1080/00273171.2015.1100528

Nichols, T. E., and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15, 1–25. doi: 10.1002/hbm.1058

Norman, K. A., Polyn, S. M., Detre, G. J., and Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430. doi: 10.1016/j.tics.2006.07.005

Oosterhof, N. N., Connolly, A. C., and Haxby, J. V. (2016). CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in matlab/GNU octave. *Front. Neuroinform.* 10, 27. doi: 10.3389/fninf.2016.00027

Pearson, K. (1903). I. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Philos. Trans. R. Soc. Lond. Ser. A* 200, 1–66. doi: 10.1098/rsta.1903.0001

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., et al. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cereb. Cortex* 28, 3095–3114. doi: 10.1093/cercor/bhx179

Sengupta, A., Kaule, F. R., Guntupalli, J. S., Hoffmann, M. B., Häusler, C., Stadler, J., et al. (2016). A studyforrest extension, retinotopic mapping and localization of higher visual areas. *Sci. Data* 3, 1–14. doi: 10.1038/sdata.2016.93

Skerry, A. E., and Saxe, R. (2014). A common neural code for perceived and inferred emotion. *J. Neurosci.* 34, 15997–16008. doi: 10.1523/JNEUROSCI.1676-14.2014

Smith, S. M., Beckmann, C. F., Andersson, J., Auerbach, E. J., Bijsterbosch, J., Douaud, G., et al. (2013). Resting-state fMRI in the human connectome project. *Neuroimage* 80, 144–168. doi: 10.1016/j.neuroimage.2013.05.039

Treder, M. S. (2020). MVPA-light: a classification and regression toolbox for multi-dimensional data. *Front. Neurosci.* 14, 289. doi: 10.3389/fnins.2020.00289

Whitfield-Gabrieli, S., and Nieto-Castanon, A. (2012). Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect.* 2, 125–141. doi: 10.1089/brain.2012.0073

Woolrich, M. W., Ripley, B. D., Brady, M., and Smith, S. M. (2001). Temporal autocorrelation in univariate linear modeling of fMRI data. *Neuroimage* 14, 1370–1386. doi: 10.1006/nimg.2001.0931

Zhang, Q. -S., and Zhu, S. -C. (2018). Visual interpretability for deep learning: a survey. *Front. Infm. Technol. Elect. Engg.* 19, 27–39. doi: 10.1631/FITEE.1700808

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

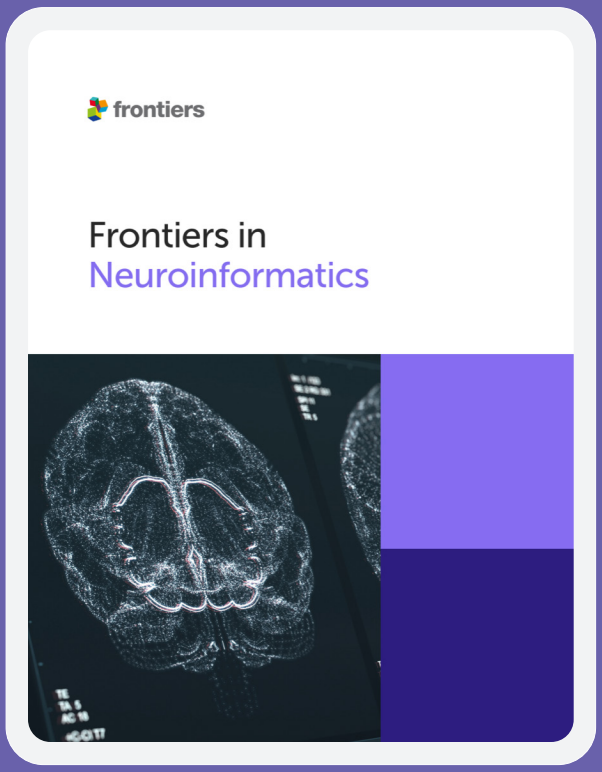
Frontiers in Neuroinformatics

Leading journal supporting neuroscience in the information age

Part of the most cited neuroscience journal series, developing computational models and analytical tools used to share, integrate and analyze experimental data about the nervous system functions.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact