Characterizing Dynamic Functional Connectivity Subnetwork Contributions in Narrative Classification with Shapley Values

Aurora Rossi, Yanis Aeschlimann, Samuel Deslauriers-Gauthier, Peter Ford Dominey

Characterizing Dynamic Functional Connectivity Subnetwork Contributions in Narrative Classification with Shapley Values

Aurora Rossi[1,*] **Yanis Aeschlimann**[2]
Samuel Deslauriers-Gauthier[2,†] **Peter Ford Dominey**[3,4,†]
[1] COATI, Université Côte d'Azur, INRIA, CNRS, I3S, Sophia Antipolis, France
[2] CRONOS, Inria Centre at Université Côte d'Azur, Sophia Antipolis, France
[3] INSERM UMR1093-CAPS, Université Bourgogne Franche-Comté,
UFR des Sciences du Sport, Dijon, France
[4] Robot Cognition Laboratory, Marey Institute Dijon, France
*Corresponding author: aurora.rossi@inria.fr
†Equal contribution

Abstract

Functional connectivity derived from functional Magnetic Resonance Imaging (fMRI) data has been increasingly used to study brain activity. In this study, we model brain dynamic functional connectivity during narrative tasks as a temporal brain network and employ a machine learning model to classify in a supervised setting the modality (audio, movie), the content (airport, restaurant situations) of narratives, and both combined. Leveraging Shapley values, we analyze subnetwork contributions within Yeo parcellations (7- and 17-subnetworks) to explore their involvement in narrative modality and comprehension. This work represents the first application of this approach to functional aspects of the brain, validated by existing literature, and provides novel insights at the whole-brain level. Our findings suggest that schematic representations in narratives may not depend solely on pre-existing knowledge of the top-down process to guide perception and understanding, but may also emerge from a bottom-up process driven by the ventral attention subnetwork.

1 Introduction

Understanding the principles of representation and computation in the human brain, and developing corresponding predictive models, remains one of the great open challenges in neuroscience. fMRI provides a rich window into the dynamics of the whole human brain with a certain level of spatial and temporal resolution. From the beginning, human language processing has been a target of investigation with fMRI [21]. Experiments with words and sentences allowed the identification of language processing areas and networks at different levels of structure [13]. More recently, evidence has emerged that language processing involves even broader recruitment across the brain, which might be obscured by time averaging and thresholding [1]. This is consistent with studies that revealed how language recruits an extended fronto-temporo-parietal semantic system beyond the classic perisylvian language network [33, 11, 5]. This has been demonstrated in the processing of narrative, full stories, which produce wide recruitment of whole brain networks for memory, visuospatial representation, and emotion [33, 12, 26]. Thus, narrative processing is a privileged context for the investigation of brain functional dynamics [32]. How can these functional dynamics be characterized? Analysis methods based on time averaging and subtraction tend to ignore the contribution of brain systems whose activity is variable and averaged out during thresholding. Functional connectivity analysis can

Preprint. Under review.

be used to capture and characterize these dynamic interactions of brain regions over time [28, 20]. Temporal brain networks model the evolution of functional connectivity over time and thus have the desired properties of capturing the full brain dynamics that may be lost in time averaging and thresholding. Here, we exploit the representational richness of dynamic functional connectivity in temporal brain networks to characterize brain dynamics during narrative processing using machine learning.

In particular, we propose a simple machine learning model to classify in a supervised setting fMRI data collected during a narrative comprehension task. The model is mainly composed of a convolutional layer and a multi-layer perceptron (MLP). It is trained to classify the modality of the narrative (audio or video), the content of the narrative (airport or restaurant situations) and these two together in a four-class classification. We use the model to investigate the importance of temporal dynamics in narrative processing and combined with the powerful explainability technique of Shapley values we delve deeper into the model's decision-making process. Specifically, we quantify the subnetwork contributions in the classification of two different parcellation methods (Yeo 7-subnetwork and 17-subnetwork) and this allows us to identify the most involved subnetworks in the narrative processing task. Our work is the first to apply this approach to functional aspects of the brain, validated by existing literature, and provides novel insights at the whole-brain level.

The results provide valuable insights, validated by existing research on narrative comprehension [3, 27], and contribute to a broader understanding of how we process narratives. Our findings challenge the initial assumption that narrative comprehension relies solely on top-down activation of scripts, where prior knowledge, experiences, and expectations solely guide interpretation [7]. The prominent role of the ventral attention subnetwork in content classification suggests a more nuanced model. This network is associated with bottom-up attentional control, implying that narrative processing might involve the assembly and integration of sensory information from the environment alongside top-down influences. This possibility aligns with the notion that schematic representations may not solely be driven by top-down activation but could be built upon bottom-up processing mediated by the ventral attention subnetwork [31].

2 Related works

Classification of tasks from fMRI data Numerous studies have explored classifying tasks and subject characteristics (such as age and sex) from functional brain connectivity data using fMRI, primarily aiming to develop powerful architectures. Examples include the work by Kim et al. [14], where they propose a Spatio-Temporal Attention Graph Isomorphism Network model for high-accuracy prediction of 7 tasks (memory, social, relational, motor, language, gambling, and emotion) alongside sex. Another approach by Kim et al. [15] utilizes a transformer to classify age, sex, and cognitive intelligence, with an integrated gradient technique for interpreting sex classification results. The latter explainability technique is also employed in a parallel similar work by Ryali et al. [22], where they classify sex using a simpler spatio-temporal deep neural network. Other papers by Huang et al. [9] and Saeidi et al. [23] use a deep learning model, mainly composed of a convolutional neural network and a recurrent neural network, and a graph neural network, respectively, to classify the 7 tasks.

Narratives classification In contrast to the aforementioned papers, our work focuses on a more detailed classification domain, specifically the classification of modalities (movie, story) and the thematic content of the script (airport, restaurant). Baldassano et al. [3] exemplify this approach, using a stochastic Hidden Markov Model to classify, based on the activation of a selection of regions of interest (ROIs) in the default attention networks, thematic content while also incorporating event alignment.

Shapley values in brain networks The use of Shapley values has become a popular approach to explain the predictions of machine learning models. In neuroscience, for instance, Amoroso et al. [2] classify three conditions (Alzheimer's disease, mild cognitive impairment, and healthy controls) based on brain structural connectivity data from MRI scans. They then leverage Shapley values to identify the most influential "patch" for classification. Another study by Kotter et al. utilizes Shapley ratings in macaque brain networks, employing a graph theory approach to analyze these networks. Here, the number of strongly connected components within a subgraph serves as the Shapley value

function [16]. The most similar work to ours is by Li et al. [17]. They propose a new estimation method for Shapley values and apply it when classifying functional connectivity data from fMRI. In their example, they classify patient conditions (autism spectrum disorder or healthy) and compute the importance of different ROIs in classification, though they don't delve into the neuroscientific interpretation of the results.

3 Our contribution

This study combines machine learning with explainable AI to investigate the specific roles of brain subnetworks during tasks involving narratives. We leverage functional connectivity, extracted from fMRI data, and Shapley values to identify which brain subnetworks are most influential in classifying narrative modality (audio and movie), thematic content (airport and restaurant situation) and their combination. The fMRI data are segmented into 7 or 17 Yeo subnetworks using the Schaefer 100 element parcellation [24]. Our machine learning model, composed of a convolutional neural network and multi-layer perceptron, achieves high accuracy and reveals the specific contributions of Yeo subnetworks in narrative processing. Importantly, the focus of our analysis is functional connectivity, rather than activation. This analysis, validated by neuroscientific interpretation aligned with existing literature, offers new insights into the functional roles of these subnetworks and the factor of time during narrative classification. Our work demonstrates the power of explainable AI in unveiling the complex interplay between brain activity and narrative comprehension. It not only helps to understand narrative processing but also paves the way for applying this approach to other areas of brain research.

4 Model

Our model takes as input a temporal brain network. This network is a sequence of brain networks, each reflecting the brain's functional connectivity at a specific time step (further details regarding the data processing are provided in the Experiments section). Mathematically, the temporal brain network can be represented as a three-dimensional tensor, denoted by $X \in [-1,1]^{R\times R\times T}$, where R represents the number of brain regions and T represents the number of time steps. In our case, R is 100 and T is 8.

The model architecture consists of a single-layer three-dimensional convolutional neural network, followed by a max pooling layer and a multi-layer perceptron for classification. The convolution filter has size (R, R, τ) with no padding, where the two first dimensions match with those of the input. This design focuses on capturing temporal features within the brain network by restricting filter movement to the temporal axis. Max pooling is then applied to reduce the dimensionality of the extracted features. Finally, a multi-layer perceptron performs the classification task. A visual representation of the model architecture is provided in Figure 1.

Notably, when the filter size in the temporal dimension is set to 1 ($\tau = 1$), the model becomes invariant to the specific order of time steps in the input data. An analysis of the model's performance with different filter sizes is provided in the Appendix section.

Formally, given an input tensor $X \in [-1,1]^{R\times R\times T}$, the output of the convolutional layer is defined as

$$Y_{k,c} = \sigma(X * W + b)_{k,c} = \sigma(\sum_{i=1}^{R}\sum_{j=1}^{R}\sum_{p=1}^{\tau} X_{i,j,k+p-1} \cdot W_{i,j,p,c} + b_{k,c})$$

where $Y \in \mathbb{R}^{K\times C}$ is the output tensor, $W \in \mathbb{R}^{R\times R\times \tau\times C}$ is the learnable filter tensor, $b \in \mathbb{R}^{K\times C}$ is the bias matrix and C is the number of output channels. The operations $\cdot, +$ and σ, which represents the $\text{ReLU}(x) = \max(\{0, x\})$ activation function, are applied component-wise. The output tensor is then passed through a max pooling layer so that the output vector $Z \in \mathbb{R}^C$ is defined as

$$Z = \max_k Y[k, c].$$

Finally, the output passed through a multi-layer perceptron of three fully connected layers with ReLU activation functions. A fully connected layer can be defined as $V = \sigma(W \cdot Z + b)$ where V is the output of the fully connected layer, W is the weight matrix, and b is the bias vector.

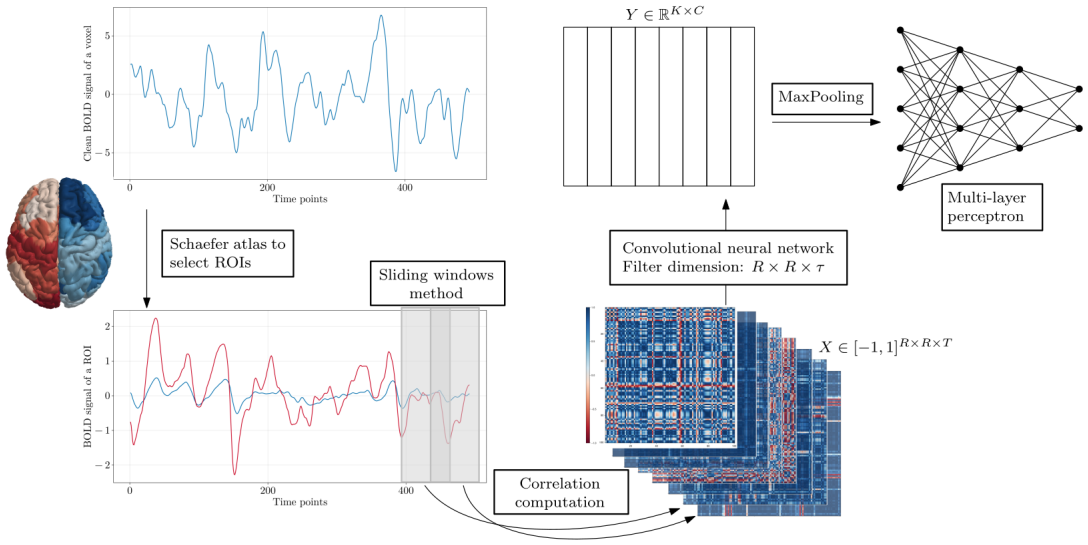


Figure 1: Pipeline from the extraction of temporal brain networks to the classification of the narrative aspects. The first step is the division of the brain into regions according to an atlas. The second step is the sliding window method, which individuates rectangular windows within which the Pearson correlation coefficient is computed between each pair of brain region time series. The output is then fed into the model, which consists of a convolutional layer, a max-pooling layer, and a multi-layer perceptron.

5 Shapley Values

Shapley values were introduced by Lloyd Shapley in 1951 in the context of cooperative game theory [25]. They quantify the contribution of each player in a coalition game. Recently, they have been adopted in machine learning to explain the predictions of models. Shapley values can be calculated using different methods including sampling or exact computation for smaller player sets [18]. In our case, we leverage Shapley values to understand the influence of specific brain subnetworks on the prediction of our model. Because of the limited number of brain subnetworks defined by the 7 Yeo parcellation method [29], we can compute the exact Shapley values. The exact Shapley value of a brain subnetwork i is defined as

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (v(S \cup \{i\}) - v(S)) \tag{1}$$

where N is the set of brain subnetworks, v is the accuracy of our model when considering the set S of brain subnetworks. To isolate the brain subnetworks in the temporal brain network X we set the entries of the other subnetworks to zero. The Shapley value $\phi_i(v)$ is the average marginal contribution of the brain subnetwork i over all possible combinations of brain subnetworks, the higher the Shapley value, the more important the brain subnetwork is for the prediction of the model. For the 17 Yeo subnetwork parcellation, the exact computation of Shapley values becomes computationally expensive. Therefore, we employ a sampling method that approximates the Shapley values using the same formula but instead of summing over all possible subnetwork combinations, we sample a large number of combinations (100 samples in our case) to approximate the average marginal contribution.

6 Experiments

Experiments were performed to determine if the temporal brain networks can be used to discriminate brain functional connectivity patterns in response to audio vs. movie narratives, airport vs. restaurant situations, and the combination of these two dimensions. We trained a machine learning model in a supervised setting to classify these aspects and used Shapley values to interpret the model's decisions.

6.1 Data

Dataset Our analysis used fMRI data from the study of Baldassano [3] archived as part of the Narratives dataset created by Nastase et al. (`https://openneuro.org/datasets/ds002345/versions/1.1.4`) [19]. The Baldassano dataset includes brain activity recordings from 31 participants engaged in a narrative task. In this task, each subject is exposed to 16 3-minute stories (4 per run over 4 runs), from two different scripts (eating at a restaurant or going through the airport). While the stories within each category share a similar high-level sequence of events, there are variations in the specific details of these events. Each run presents 2 movies and 2 audio stories, for a total of 8 movies and 8 audio segments over the course of the experiment. The dataset is balanced in terms of the number of samples per modality and content.

Preprocessing The fMRI data has a spatial resolution of $91 \times 91 \times 109$ voxels in the x, y, and z axes, respectively, for a total of $902,629$ voxels. Each voxel measures $2 \times 2 \times 2$ mm. The repetition time is 1.5 seconds, for a total of 490 time points and a total duration of 12 minutes per run approximatively.

Preprocessing involved transforming the blood-oxygen-level-dependent (BOLD) signals from each voxel into temporal graphs. We implemented a pipeline to reduce motion artifacts by performing linear regression on the movement parameters. Additionally, a bandpass filter ($0.01 - 0.08$ Hz) was applied to remove noise arising from respiration and cardiac pulsations [30].

To define the network nodes, we employed the Schaefer et al. brain atlas (after having put the data in the MNI152 space), parcellating the brain into 100 ROIs based on anatomical and functional criteria [24]. ROIs were created by averaging the BOLD time series of voxels within gray matter regions. We then utilized a sliding window approach with 30-second windows and 7.5 seconds overlap to divide the data into time steps. The Pearson correlation coefficient was computed between each pair of ROI time series within each window, with the resulting correlation value assigned as the weight of the edge connecting the corresponding ROI nodes. This process yielded an adjacency matrix for each time window, and the sequence of these matrices formed the temporal brain networks (see Figure 1).

6.2 Experimental setting

The experiments were conducted on a workstation equipped with a single NVIDIA Quadro RTX 8000 graphics card. We utilized the Julia programming language for the workflow, from network creation starting from the clean signal to the model development [4]. The Flux.jl library was used for neural network implementation and the Makie.jl library was used for visualization [10, 6]. The source code is available at the GitHub repository `https://github.com/aurorarossi/fMRINarrativeClassification`.

Hyperparameters The hyperparameters were chosen based on empirical observations. The convolutional filter τ parameter was set to 4 for the modality classification task and 8 for the content and the combined classification task (see the Appendix for more details). The number of output channels was set to 128 for all the tasks. The MLP had two hidden layers with 64 and 32 units each with a ReLU activation function. The output dimension of the MLP was set to 2 for the modality classification task, 2 for the content classification task, and 4 for the combined classification task.

Training Given the limited size of the dataset, we employed a batch size of 1 during training. We used the Adam optimizer with a learning rate of 0.0001. The training process lasted for 20 epochs. The choice of 20 epochs was determined through experiments to achieve a good balance between training time and model performance. For the loss function, we used either logit binary cross-entropy or logit cross-entropy depending on the number of classes in the task. To ensure robustness against potential variations due to model initialization, we retrain the model 15 times with different random splits of the data (80% training, 20% testing). During each iteration, we compute both the Shapley values and the model's accuracy. Finally, we report the mean and standard deviation to account for variability for the accuracy, and for Shapley values of each subnetwork, we present the mean values along with error bars representing the standard deviation. This approach ensures a comprehensive understanding of the model's performance, the contribution of individual brain subnetworks to its classifications, and the robustness of these findings across model initializations.

7 Results

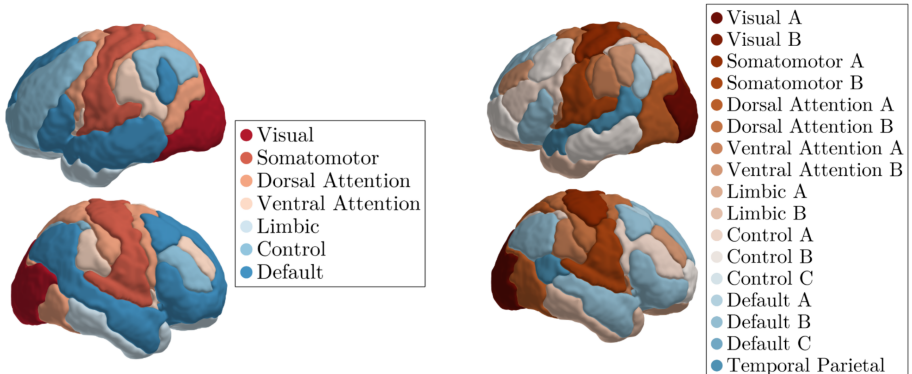
In this section, we describe the results of our experiments. We present the performance of the model on three classification tasks:

- **Modality classification**: this task focuses on classifying the brain network based on the modality of the stimuli, audio or movie.
- **Content classification**: the model classifies the brain network based on the content of the stimuli, airport or restaurant situations.
- **Combined Modality and Content Classification**: this task evaluates the model's ability to jointly classify both the modality and the content of the stimuli.

	Modality	Content	Both Modality and Content
Accuracy	$96.32\% \pm 1.36\%$	$80.9\% \pm 1.75\%$	$80.70\% \pm 2.97\%$
Precision	$95.64\% \pm 1.43\%$	$84.55\% \pm 2.29\%$	$81.54\% \pm 5.34\%$
Recall	$97.08\% \pm 2.20\%$	$75.69\% \pm 3.02\%$	$80.70\% \pm 5.23\%$
F1-Score	$96.34\% \pm 1.36\%$	$79.84\% \pm 1.96\%$	$80.92\% \pm 6.06\%$
Accuracy permuting times	$86.60\% \pm 3.36\%$	$63.19\% \pm 4.40\%$	$53.12\% \pm 5.85\%$

Table 1: Performance metrics of the model on the modality, content, and combined classification tasks. The last row shows the model's performance when the time steps of the brain networks are permuted.

The results in Table 1 show that the model performs well on the modality classification task, achieving an accuracy of $96.32\% \pm 1.36\%$. While still a good performance considering the complexity, the model's accuracy on the content classification task was slightly lower at $80.9\% \pm 1.75\%$. This difference might be attributed to the inherent difficulty of content classification compared to modality identification. Furthermore, the combined modality and content classification task resulted in an accuracy of $80.70\% \pm 2.97\%$, which is consistent with the content classification task. Notably, the model displayed consistent performance across all metrics.



(a) (b)

Figure 2: Yeo parcellations used in the Shapley value analysis. The 7-subnetwork parcellation is shown on the left (a), while the 17-subnetwork parcellation is shown on the right (b).

To assess the importance of the time dimension in the classification tasks, we permuted the time steps of the brain networks and retrained the model. The results in the last row of Table 1 demonstrate a significant drop in accuracy: 10% for modality classification, 17% for content classification,

and a substantial 27% for the combined task. These drops strongly suggest that the temporal dynamics of the brain networks play a crucial role in all classification tasks and that the model leverages this information effectively. The performance decrease is more pronounced in content and combined classification tasks compared to modality classification. This aligns with our expectations. Understanding content, which often unfolds over time and involves complex relationships between brain regions, is likely more dependent on the temporal dynamics of brain activity compared to simply identifying the modality.

To gain deeper insights into how the model leverages brain activity for classification, we employed Shapley values. Here, we focus on subnetworks defined by the Yeo parcellation method [29], specifically the 7-subnetwork and 17-subnetwork parcellations. Visualizations of these parcellations are provided in Figure 2. Black and white compatible versions of these figures can be found in the Appendix.

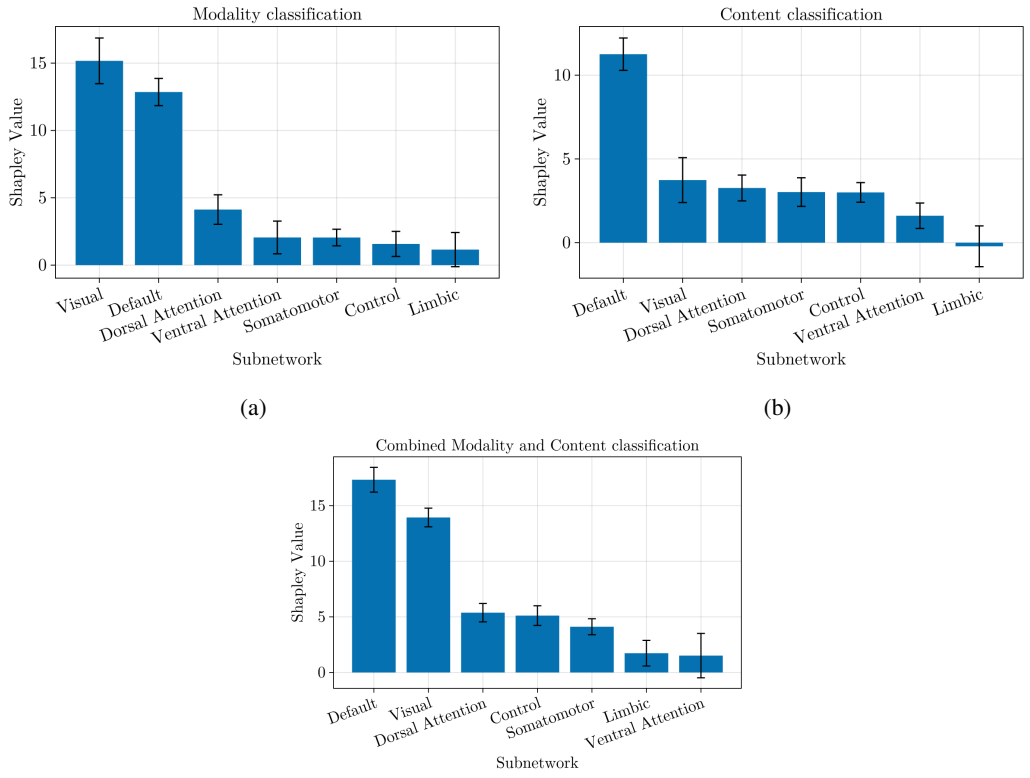


Figure 3: This figure shows the contribution of Yeo 7-subnetworks computed with Shapley values for classifying narrative using a machine learning model. The bars represent the average contribution of each subnetwork to the model's predictions, with higher values indicating greater influence. The error bars represent the standard deviation of the Shapley values.

Figure 3 presents the Shapley values for the 7-subnetwork parcellation. In the modality classification task, the visual subnetwork emerges as the most influential, followed by the default mode subnetwork (Figure 3a). This aligns with the intuitive notion that processing visual information plays a key role in distinguishing modalities. For the content classification task, the high value of the default mode subnetwork suggests its influence in understanding the meaning and content of the stimuli as suggested by previous studies [3, 27] (Figure 3b). Finally, the combined classification task reveals the importance of both the visual and default mode networks (Figure 3c), suggesting that the model utilizes a combination of visual features and higher-order processing for accurate content and modality classification.

Figure 4 presents the Shapley values for the 17-subnetwork parcellation. In the modality classification task, the visual A and B, default A and B and somatomotor A subnetworks emerge as the most

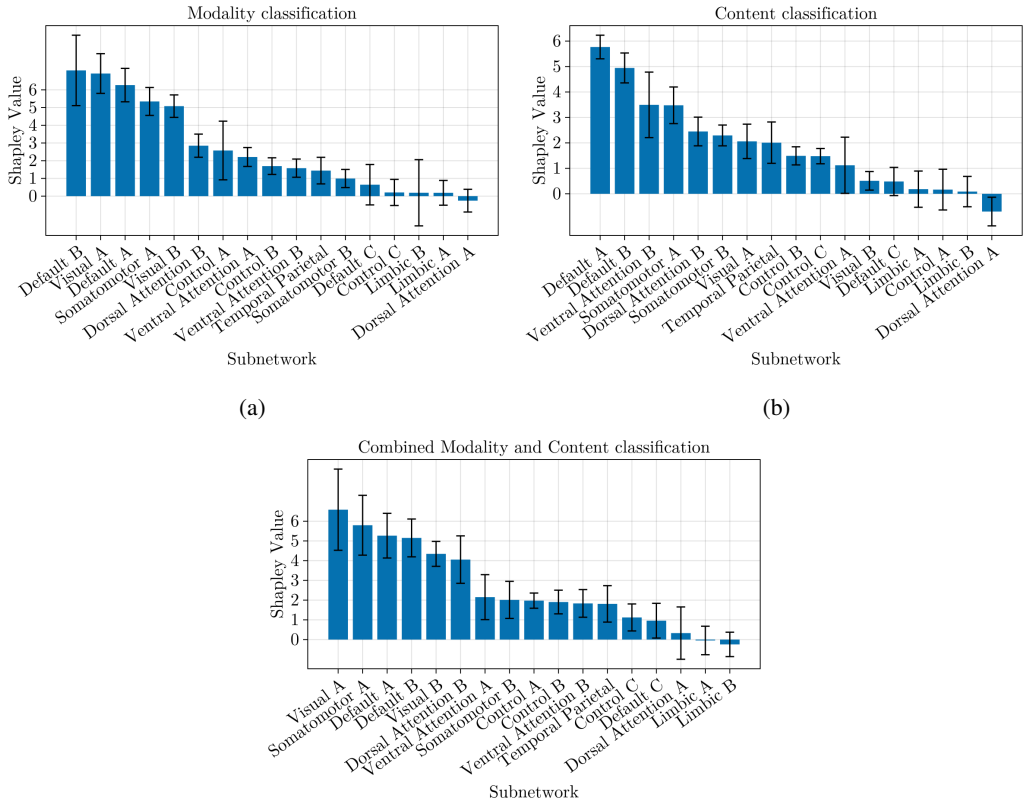


Figure 4: This figure shows the contribution of Yeo 7-subnetworks computed with Shapley values for classifying narrative using a machine learning model. The bars represent the average contribution of each subnetwork to the model's predictions, with higher values indicating greater influence. The error bars represent the standard deviation of the Shapley values.

influential (Figure 4a). For the content classification task, the default A and B subnetworks, the somatomotor A and the ventral attention B also play crucial roles (Figure 4b). Finally, the combined classification task reveals the importance of the visual A and B, default A and B, and somatomotor A subnetworks (Figure 4c).

8 Discussion

This work investigated the neural basis of narrative processing using a machine learning model that classifies narrative aspects (modality, content, combined) based on functional connectivity networks derived from fMRI data. The model's performance aligned with expectations: higher accuracy for modality classification, which is a simpler task because it relies on sensory information, compared to content classification which requires a deeper understanding of the narrative. Permuting time steps in the temporal brain network significantly reduced accuracy, particularly in content and combined tasks, suggesting that temporal dynamics rely on the sequence of events to understand the content.

To delve deeper into the model's decision-making process, we employed Shapley values, a powerful explainable AI technique that quantifies subnetwork contributions. The results provide insights into the importance of brain subnetworks of two different parcellation methods (Yeo 7-subnetwork and 17-subnetwork) during narrative comprehension and contribute to the broader understanding of how the brain processes narratives.

Our findings revealed that in the 7-subnetwork analysis, the visual and default subnetworks are key for modality classification, reflecting the intuitive notion that visual processing is essential

for distinguishing between movies and audio stories. In content classification, the default mode subnetwork emerged as the most influential, suggesting its essential function in understanding the meaning and content of the stimuli. This aligns with existing research that has highlighted the default mode subnetwork's involvement in higher-order cognitive functions, such as narrative comprehension [3, 27]. The combined classification task emphasized the importance of both visual and default mode networks, as expected.

A more fine-grained analysis using the 17-subnetwork parcellation revealed additional insights. While visual and default mode networks remained dominant for modality classification, the somatomotor subnetwork also showed a high Shapley value. The latter can be better understood in the context of embodied cognition and language comprehension. A seminal study of embodied language comprehension demonstrated that passive reading of action words produces a corresponding somatotopic activation of the motor and premotor cortex [8]. Likewise, viewing images or reading sentences describing everyday actions produces a distributed activation in fronto-temporo-parietal network that includes sensory-motor and premotor cortex [11]. Similar to the 7-subnetwork analysis, the default mode subnetwork was most influential for content classification. Interestingly, the ventral attention subnetwork also played a significant role. This finding is a step further to answer the open question raised by the Baldassano et al. 2018 study [3]. They proposed that schematic representations in the brain might not solely rely on top-down activation of scripts in the medial prefrontal cortex. They suggested these representations could serve as building blocks for a complete narrative script formed through a bottom-up process. Our observation of a high Shapley value for the ventral attention subnetwork, which is known to be also associated with bottom-up attentional control, aligns with this possibility. Finally, the combined classification task again highlighted the importance of visual, default mode, and somatomotor A networks.

Limitations and Future Works It is important to acknowledge that the primary limitation of this study is the size of the dataset used. This may limit the generalizability of our findings to other populations or narrative stimuli. Future research could address this by employing larger datasets, if available. Additionally, exploring the generalizability of these findings across diverse datasets would be valuable. Within the context of the current dataset size, future work could delve deeper into other aspects of narrative processing. One potential direction is to investigate the impact of individual differences in narrative comprehension. For instance, research could explore how factors such as age, reading experience, or cultural background might influence how individuals process narratives based on brain network activity. Another promising avenue for future research involves investigating the impact of different parcellation methods on the results of Shapley values. Currently, we employ the Yeo 7-subnetwork and 17-subnetwork parcellations. However, more fine-grained parcellations might reveal even more nuanced insights into the specific subnetworks involved in narrative processing. Exploring these possibilities could lead to a more comprehensive understanding of the role of brain regions in narrative comprehension.

Conclusion Overall, our work demonstrates the potential of combining machine learning models with explainable AI techniques like Shapley values to understand the role of brain subnetworks during narrative processing. Our findings not only contribute to a deeper understanding of how the brain processes narratives but also showcase the broader applicability of this approach. In tasks where the role of specific brain regions remains unclear, this methodology can provide valuable new insights. By highlighting subnetwork contributions through Shapley values, we can generate novel hypotheses about the functional roles of these regions. In our case, the model's performance aligns with existing literature on narrative comprehension, validating the approach. Importantly, this research validates an alternative and complementary method for investigating brain function in human cognition, which involves functional connectivity. This successful validation paves the way for further exploration of brain networks not only in higher-order cognition, motor tasks, and emotional processing but also in any domain where the neural basis remains partially understood.

Acknowledgements

A.R. would like to thank Emanuele Natale for the fruitful discussions and his help in designing the model. A.R. would also like to thank Pierluigi Crescenzi for the discussion on the explainability technique used in the paper. This work has been supported by the French government, through the

UCA DS4H Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-17-EURE-0004 and the ANR France Relance project.

Author contributions

A.R. processed the data, designed the model, performed the experiments, wrote the original draft and edited the manuscript. Y.A. preprocessed the data and reviewed and edited the manuscript. S.D.G. and P.F.D. conceived the project, supervised the project, and reviewed and edited the manuscript.

References

[1] Sarah Aliko, Bangjie Wang, Steven L Small, and Jeremy I Skipper. *The entire brain, more or less, is at work* : 'Language regions' are artefacts of averaging, September 2023.

[2] Nicola Amoroso, Silvano Quarto, Marianna La Rocca, Sabina Tangaro, Alfonso Monaco, and Roberto Bellotti. An eXplainability Artificial Intelligence approach to brain connectivity in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 15, August 2023.

[3] Christopher Baldassano, Uri Hasson, and Kenneth A. Norman. Representation of Real-World Event Schemas during Narrative Perception. *The Journal of Neuroscience*, 38(45):9689–9699, November 2018.

[4] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[5] Jeffrey R. Binder and Rutvik H. Desai. The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11):527–536, November 2011.

[6] Simon Danisch and Julius Krumbiegel. Makie.jl: Flexible high-performance data visualization for Julia. *Journal of Open Source Software*, 6(65):3349, 2021.

[7] Fraida Dubin and David Bycina. Academic reading and the esl/efl teacher. *Teaching English as a second or foreign language*, 2:195–215, 1991.

[8] Olaf Hauk, Ingrid Johnsrude, and Friedemann Pulvermüller. Somatotopic Representation of Action Words in Human Motor and Premotor Cortex. *Neuron*, 41(2):301–307, January 2004.

[9] Xiaojie Huang, Jun Xiao, and Chao Wu. Design of Deep Learning Model for Task-Evoked fMRI Data Classification. *Computational Intelligence and Neuroscience*, 2021:1–10, August 2021.

[10] Michael Innes, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, Neethu Mariya Joy, Tejan Karmali, Avik Pal, and Viral Shah. Fashionable modelling with flux. *CoRR*, abs/1811.01457, 2018.

[11] A.L. Jouen, T.M. Ellmore, C.J. Madden, C. Pallier, P.F. Dominey, and J. Ventre-Dominey. Beyond the word and image: characteristics of a common meaning system for language and vision revealed by functional and structural imaging. *NeuroImage*, 106:72–85, February 2015.

[12] Iiro P. Jääskeläinen, Mikko Sams, Enrico Glerean, and Jyrki Ahveninen. Movies and narratives as naturalistic stimuli in neuroimaging. *NeuroImage*, 224:117445, 2021.

[13] Timothy A. Keller, Patricia A. Carpenter, and Marcel Adam Just. The Neural Bases of Sentence Comprehension: a fMRI Examination of Syntactic and Lexical Processing. *Cerebral Cortex*, 11(3):223–237, 03 2001.

[14] Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. In *Advances in Neural Information Processing Systems*, volume 34, pages 4314–4327. Curran Associates, Inc., 2021.

[15] Peter Kim, Junbeom Kwon, Sunghwan Joo, Sangyoon Bae, Donggyu Lee, Yoonho Jung, Shinjae Yoo, Jiook Cha, and Taesup Moon. Swift: Swin 4d fmri transformer. In *Advances in Neural Information Processing Systems*, volume 36, pages 42015–42037. Curran Associates, Inc., 2023.

[16] Rolf Kötter. Shapley ratings in brain networks. *Frontiers in Neuroinformatics*, 1, 2007.

[17] Xiaoxiao Li, Yuan Zhou, Nicha C. Dvornek, Yufeng Gu, Pamela Ventola, and James S. Duncan. Efficient Shapley Explanation for Features Importance Estimation Under Uncertainty. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12261. Springer International Publishing, 2020.

[18] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777. Curran Associates Inc., 2017.

[19] Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily Micciche, Gina Choe, Ariel Goldstein, Tamara Vanderwal, Yaroslav O. Halchenko, Kenneth A. Norman, and Uri Hasson. "narratives", 2020.

[20] Maria Giulia Preti, Thomas AW Bolton, and Dimitri Van De Ville. The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage*, 160:41–54, 2017. Functional Architecture of the Brain.

[21] Cathy J. Price. A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816–847, August 2012.

[22] Srikanth Ryali, Yuan Zhang, Carlo de Los Angeles, Kaustubh Supekar, and Vinod Menon. Deep learning models reveal replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization. *Proceedings of the National Academy of Sciences*, 121(9):e2310012121, 2024.

[23] Maham Saeidi, Waldemar Karwowski, Farzad V. Farahani, Krzysztof Fiok, P. A. Hancock, Ben D. Sawyer, Leonardo Christov-Moore, and Pamela K. Douglas. Decoding Task-Based fMRI Data with Graph Neural Networks, Considering Individual Differences. *Brain Sciences*, 12(8):1094, August 2022.

[24] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

[25] Lloyd S Shapley. Notes on the n-person game—ii: The value of an n-person game. 1951.

[26] Lauren J. Silbert, Christopher J. Honey, Erez Simony, David Poeppel, and Uri Hasson. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, 111(43), October 2014.

[27] Erez Simony, Christopher J Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, 7(1):12141, July 2016.

[28] Ann E. Sizemore and Danielle S. Bassett. Dynamic graph metrics: Tutorial, toolbox, and tale. *NeuroImage*, 180:417–427, 2018. Brain Connectivity Dynamics.

[29] B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, September 2011.

[30] Koene R. A. Van Dijk, Trey Hedden, Archana Venkataraman, Karleyton C. Evans, Sara W. Lazar, and Randy L. Buckner. Intrinsic Functional Connectivity As a Tool For Human Connectomics: Theory, Properties, and Optimization. *Journal of Neurophysiology*, 103(1):297–321, January 2010.

[31] Simone Vossel, Joy J. Geng, and Gereon R. Fink. Dorsal and Ventral Attention Systems: Distinct Neural Circuits but Collaborative Roles. *The Neuroscientist*, 20(2):150–159, April 2014.

[32] Roel M. Willems, Samuel A. Nastase, and Branka Milivojevic. Narratives for Neuroscience. *Trends in Neurosciences*, 43(5):271–273, May 2020.

[33] Jiang Xu, Stefan Kemeny, Grace Park, Carol Frattali, and Allen Braun. Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage*, 25(3):1002–1015, April 2005.

A Appendix

A.1 Choice of parameter τ

The following figure shows the evolution of the model's accuracy as a function of the third dimension of the convolutional filter (i.e. τ). For the modality classification, we set $\tau = 4$, since model performance seems not to increase significantly beyond this value (Figure 5a). For the content and combined classification, we set $\tau = 8$m since the model performance seems the best for this value (Figure 5b and Figure 5c). It is important to highligh that when $\tau = 8$ the convolution behaviour is similar to the one of dense layer.

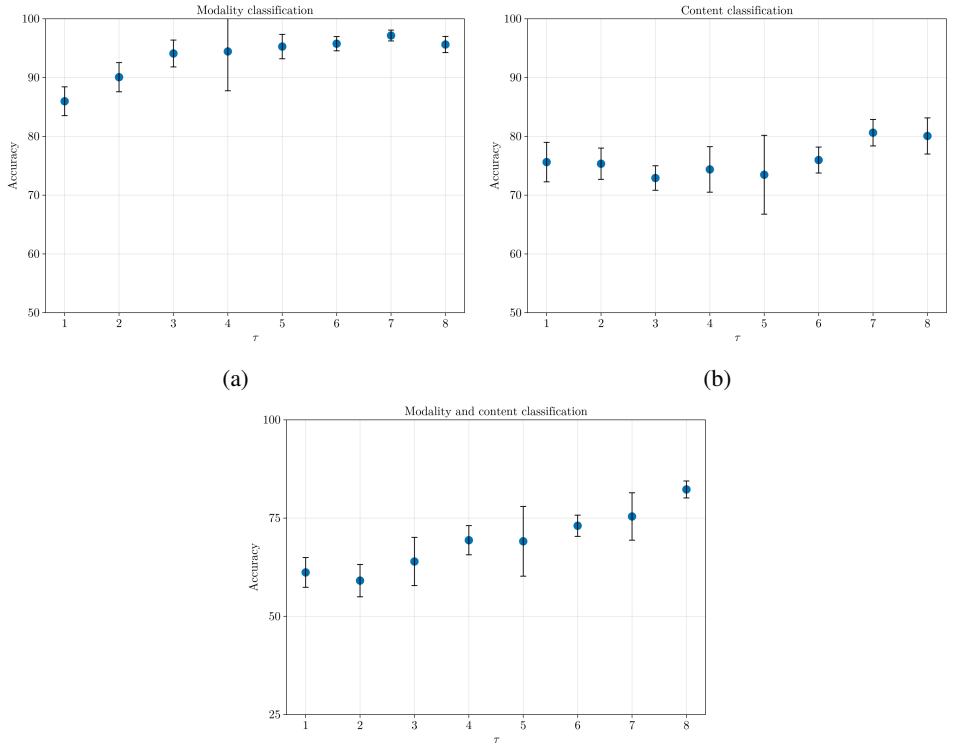


Figure 5: Model's accuracy as a function of the third dimension of the convolutional filter.

A.2 Yeo parcellations black and white compatible

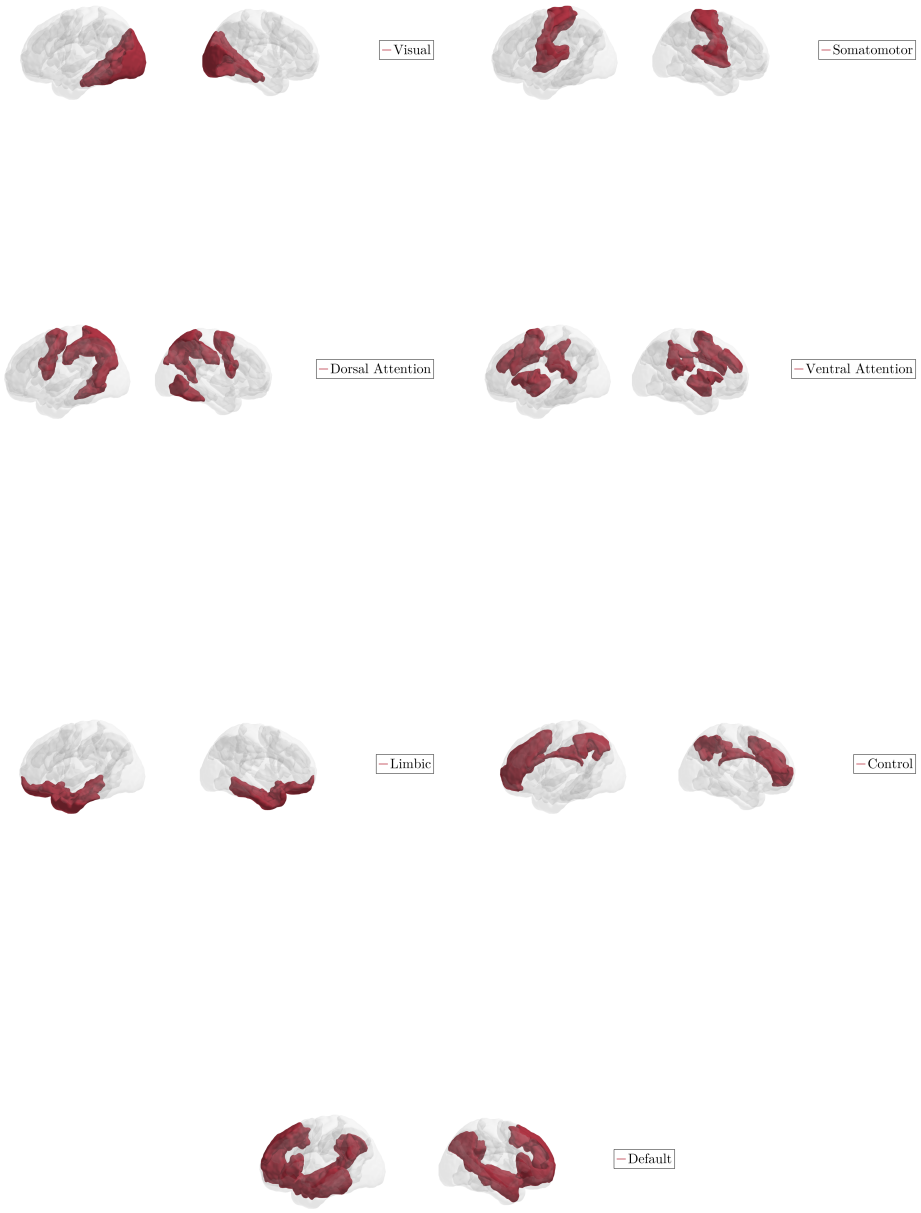
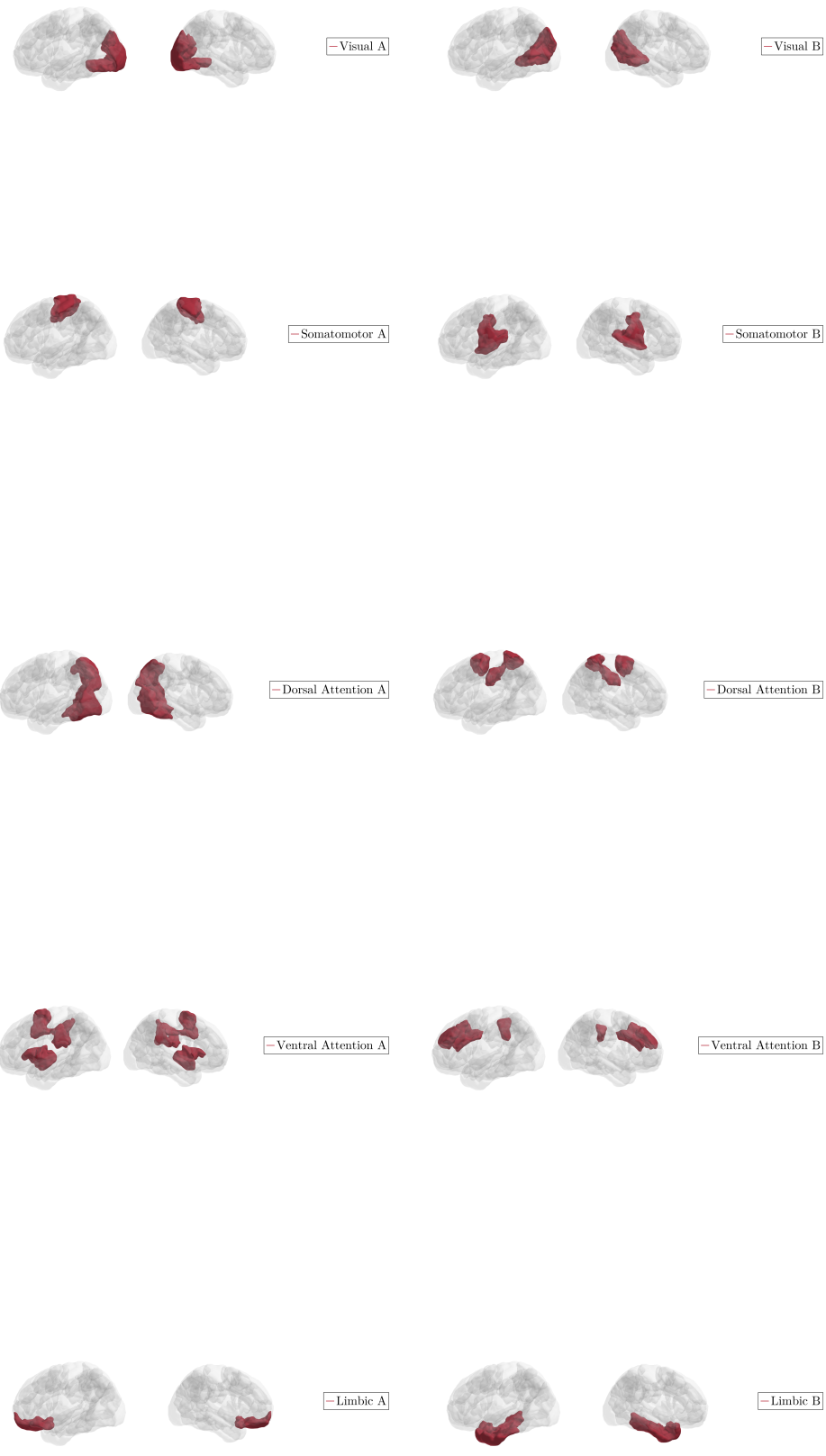


Figure 6: Yeo 7-subnetworks


Visual A
Visual B
Somatomotor A
Somatomotor B
Dorsal Attention A
Dorsal Attention B
Ventral Attention A
Ventral Attention B
Limbic A
Limbic B

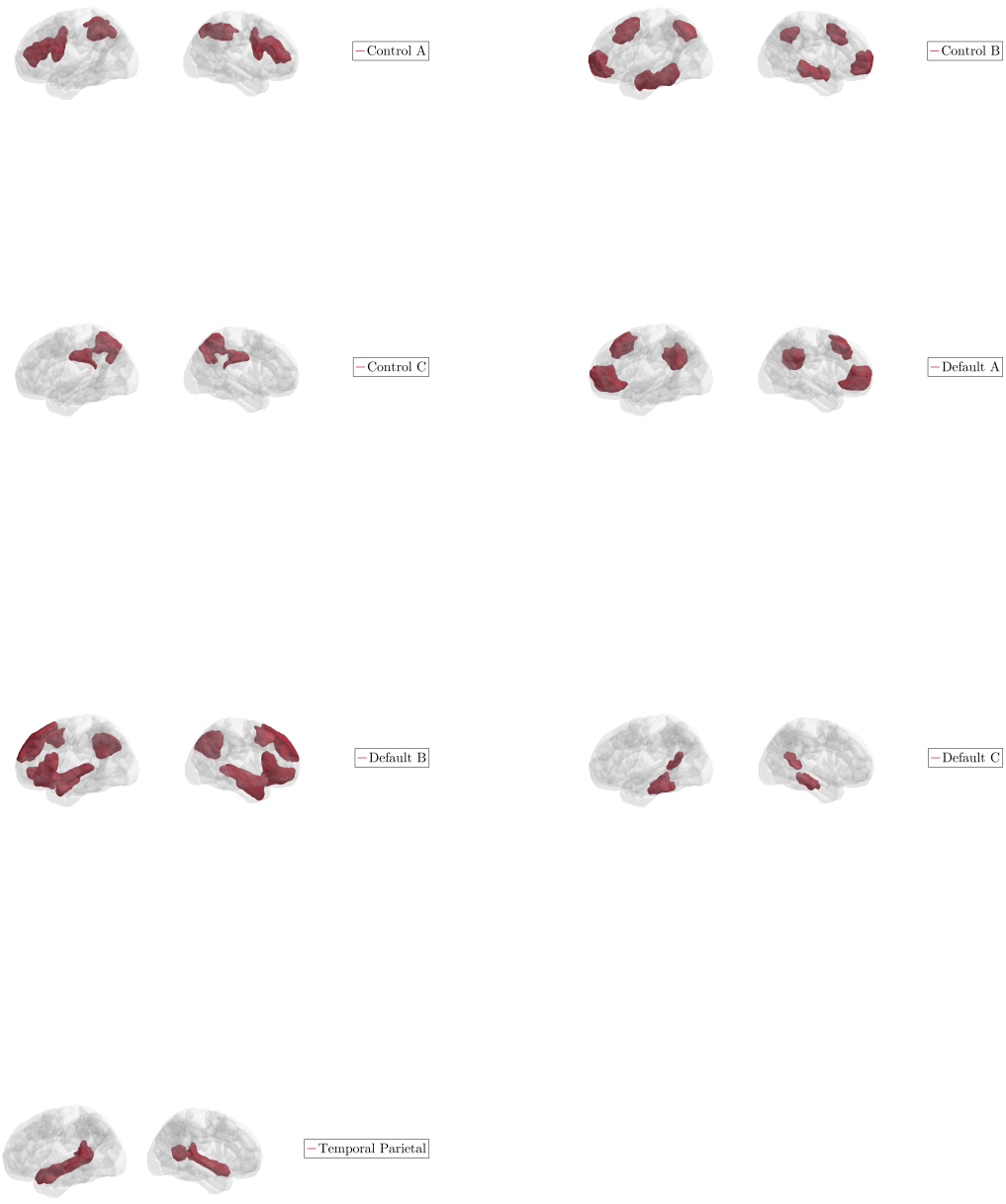


Figure 8: Yeo 17-subnetworks