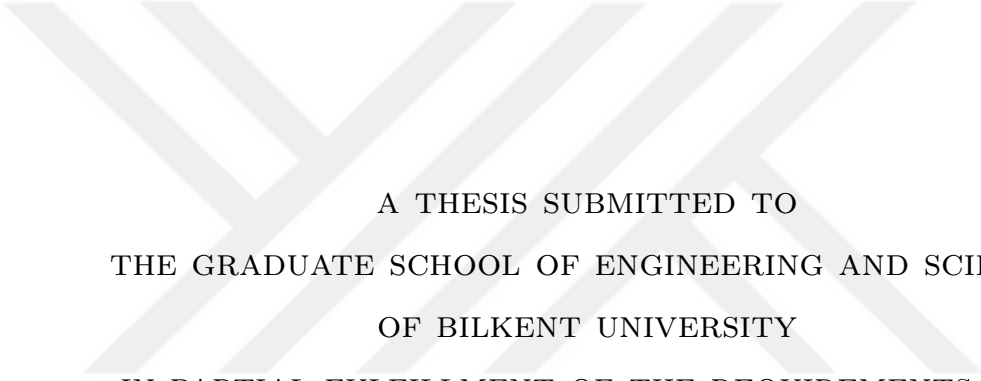


NOVEL DEEP LEARNING APPROACHES FOR FUNCTIONAL FMRI DATA ANALYSIS



A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

By
Hasan Atakan Bedel
August 2024

Novel Deep Learning Approaches for Functional FMRI Data Analysis

By Hasan Atakan Bedel

August 2024

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Tolga Çukur(Advisor)

Emine Ülkü Sarıtaş Çukur

Abdullah Aydın Alatan

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

NOVEL DEEP LEARNING APPROACHES FOR FUNCTIONAL FMRI DATA ANALYSIS

Hasan Atakan Bedel
M.S. in Electrical and Electronics Engineering
Advisor: Tolga Çukur
August 2024

Functional MRI (fMRI) has revolutionized our ability to analyze brain activity by providing insights into high-dimensional, time-series data. Despite advancements, existing methods often fall short in their ability to effectively capture contextual representations across varying time scales and interpret the resulting data. Addressing these challenges, this thesis introduces three innovative approaches: BolT, GraphCorr, and DreaMR, each designed to enhance the analysis and interpretability of fMRI data. BolT represents a significant advancement in modeling fMRI time series by utilizing a blood-oxygen-level-dependent transformer architecture. This model incorporates a novel fused window attention mechanism, which enables the extraction of both local and global representations by processing temporally-overlapped windows and employing cross-window regularization. BolT’s approach improves upon existing methods, offering enhanced sensitivity and aligning with key neuroscientific findings through extensive experimentation. GraphCorr addresses limitations in static functional connectivity (FC) features used in classification models by introducing a graph neural network-based plug-in. This method captures dynamic latent FC features while preserving dimensional efficiency, employing a node embedder and lag filter module to refine temporal information. The integration of these features through a message passing algorithm significantly enhances the performance of baseline classification models, as demonstrated through comprehensive testing on public datasets. Finally, DreaMR tackles the interpretability of deep fMRI classifiers through a novel diffusion-driven counterfactual approach. By using fractional multi-phase-distilled diffusion, DreaMR generates high-fidelity counterfactual samples and employs a transformer architecture to account for long-range spatiotemporal contexts. This method surpasses traditional counterfactual techniques in both fidelity and efficiency, offering a more precise and actionable explanation of classifier decisions. Together, these contributions advance the field of fMRI analysis

by improving model performance and interpretability, thereby facilitating more effective and insightful neuroimaging research.



Keywords: Novel Deep Learning Approaches for Functional MRI Data Analysis.

ÖZET

FONKSİYONEL MRG VERİ ANALIZI İÇİN YENİLİKÇİ DERİN ÖĞRENME YÖNTEMLERİ

Hasan Atakan Bedel

Elektrik ve Elektronik Mühendisliği, Yüksek Lisans

Tez Danışmanı: Tolga Çukur

Ağustos 2024

Fonksiyonel MRI (fMRI), beyin aktivitesini analiz etme yeteneğimizi devrim niteliğinde artırmış ve yüksek boyutlu, zaman serisi verilerine dair içgörüler sunmuştur. Gelişmelere rağmen, mevcut yöntemler genellikle farklı zaman ölçeklerinde bağlamsal temsilleri etkili bir şekilde yakalama ve elde edilen verileri yorumlama konusunda yetersiz kalmaktadır. Bu zorlukları ele alarak, bu tez üç yenilikçi yaklaşım sunmaktadır: BolT, GraphCorr ve DreaMR. Her biri fMRI verilerinin analizini ve yorumlanabilirliğini artırmayı hedeflemektedir. BolT, fMRI zaman serilerini modellemede önemli bir ilerlemeyi temsil eder ve kan-serbest oksijen seviyesine bağlı (BOLD) bir transformer mimarisi kullanır. Bu model, zamanla örtüşen pencereleri işleyerek ve pencere arası düzenleme uygulayarak hem yerel hem de küresel temsilleri çıkaran yenilikçi bir füzyonlu pencere dikkat mekanizması içerir. BolT'nin yaklaşımı, mevcut yöntemleri geliştirir, hassasiyeti artırır ve kapsamlı deneylerle önemli nörobilimsel bulgularla uyum sağlar. GraphCorr, sınıflandırma modellerinde kullanılan statik fonksiyonel bağlantı (FC) özelliklerindeki sınırlamaları ele alır ve bir grafik sinir ağı tabanlı eklenti sunar. Bu yöntem, dinamik gizli FC özelliklerini boyutsal verimliliği koruyarak yakalar ve zaman bilgilerini rafine etmek için bir düğüm gömme ve gecikme filtresi modülü kullanır. Bu özelliklerin mesaj iletimi algoritması aracılığıyla entegrasyonu, temel sınıflandırma modellerinin performansını önemli ölçüde artırır ve kamu veri setlerinde kapsamlı testlerle gösterilmiştir. Son olarak, DreaMR, derin fMRI sınıflandırıcılarının yorumlanabilirliğini yenilikçi bir difüzyon destekli karşı-faktüel yaklaşımıyla ele alır. DreaMR, fraksiyonel çok-aşamalı damıtılmış difüzyon kullanarak yüksek-fideliteli karşı-faktüel örnekler üretir ve uzun menzilli uzaysal-zamansal bağlamları hesaba katmak için bir transformer mimarisi kullanır. Bu yöntem, geleneksel karşı-faktüel teknikleri hem fidelite hem de verimlilik açısından geride bırakır ve sınıflandırıcı kararlarının daha hassas ve eyleme

geçirilebilir bir açıklamasını sunar. Bu katkılar, fMRI analizini model performansını ve yorumlanabilirliğini geliştirerek ilerletir ve daha etkili ve içgörülü nörogörüntüleme araştırmalarını kolaylaştırır.



Anahtar sözcükler: sinir bilimi, tıbbi görüntü analizi, derin öğrenme, dönüştürücü, difüzyon modelleri.

Acknowledgement

My deepest thanks to Prof. Tolga Cukur, who has been an incredible mentor and guide throughout this journey. His unwavering support, whether it was mental, helping with writing, organizing, or editing papers, or diving into deep discussions about ideas, has meant the world to me. I'm truly grateful for his guidance and encouragement every step of the way.

I want to thank every staff member at Bilkent University, from those who clean my dorm to those who keep the gym running smoothly, allowing me to benefit from it, and the refectory staff who provide meals every day, even during holidays. They all played a crucial role in helping me function each day, living in a clean space, keeping me well-fed, and helping me relax at the gym.

I also want to thank all the cats at Bilkent. My heart always warmed when I saw any of you, and in those moments, my problems seemed to dissolve. You gave me the perfect breathing break every time.

I want to thank our former lab members—Salman, Muzaffer, Onat, and Mahmut—for their invaluable mental support. They were always there to listen, inspire me, and discuss research ideas, and their positivity meant the world to me. I hope to pay it forward by offering the same support to the newcomers in our lab.

I also want to thank Irmak for sharing a part of this research journey with me and for those uplifting Zoom talks from Stanford. I'm thrilled that we'll continue to create new memories throughout the rest of our PhD journey together.

I want to thank my lovely friends, Hasan, Atakan (no, they're not me—well, maybe they are, haha), and Emirhan, for making the last year of my master's unforgettable. It's truly been a year I'll never forget. Hasan, thank you for being there for me when I was sick, when it was raining, when I needed to share my burdens, and for always smiling and spreading positive energy among us. You

were always driving us to places where we had the time of our lives. Atakan, thank you for breaking the ice and becoming a dear friend, and for helping to form this amazing friend group. I appreciate you listening and being there for me when I needed it. Emirhan, I feel like we understand each other better than most, and you really helped me through the tough times in the first year of my master's. Thank you for that, and I'm sorry for that gap in our friendship, but I think we've recovered well. You're the perfect spice that our friend group needed—and always will be. Thank you for your understanding and support throughout this journey. I love you guys, and I know our wild adventures will continue in California. I'm confident each of you will reach great heights!

Thank you, XX, for an incredibly interesting year. It's been a roller-coaster ride, full of experiences that have taught me so much about myself, and I owe much of that to you. I believe this journey has been new and intense for both of us, and I genuinely feel that I've grown a lot because of it. I know you have many funny and weird adventures ahead, with plenty more to experience. I hope you find the peace you deserve along the way.

Thank you, Mom and Dad, for always being there for me and for your unwavering support throughout my education, both emotionally and financially. You put your own needs aside for the sake of my sister's and my education. I sincerely hope that now you can focus on your own lives and enjoy the rest, pursuing your own dreams.

Contents

1	Introduction	1
1.1	Deep Learning Applications in Functional MRI Analysis	1
1.2	Contributions	3
1.3	Outline	5
2	Fundamentals of Functional Magnetic Resonance Imaging (fMRI)	6
2.1	Fundamentals of MRI	6
2.1.1	Basic Principles	6
2.1.2	Imaging Process	7
2.1.3	Mathematical Modeling	8
2.1.4	Machine Learning in MRI	8
2.2	Fundamentals of fMRI	9
2.2.1	Basic Principles	10

2.2.2	Encoding Models of fMRI	11
-------	-----------------------------------	----

2.2.3	Decoding Models of fMRI	12
-------	-----------------------------------	----

3 BolT: Fused Window Transformers for fMRI Time Series Analysis **14**

3.1	Introduction	15
-----	------------------------	----

3.2	Related Work	17
-----	------------------------	----

3.2.1	Traditional methods	17
-------	-------------------------------	----

3.2.2	Deep learning methods on FC features	18
-------	--	----

3.2.3	Deep learning methods on BOLD responses	18
-------	---	----

3.3	Theory	22
-----	------------------	----

3.3.1	BolT	22
-------	----------------	----

3.3.2	Explanatory technique	25
-------	---------------------------------	----

3.4	Methods	28
-----	-------------------	----

3.4.1	Datasets	28
-------	--------------------	----

3.4.2	Experimental procedures	29
-------	-----------------------------------	----

3.4.3	Implementation details	29
-------	----------------------------------	----

3.4.4	Model complexity	32
-------	----------------------------	----

3.4.5	Competing methods	33
-------	-----------------------------	----

3.5	Results	37
-----	-------------------	----

3.5.1	Ablation studies	37
3.5.2	Comparative demonstration of BolT	44
3.5.3	Explanability of BolT	45
3.6	Discussion	47
4	A plug-in graph neural network to boost temporal sensitivity in fMRI analysis	52
4.1	Introduction	52
4.2	Related Work	55
4.3	Methods	57
4.3.1	Graph formation	58
4.3.2	Network architecture	59
4.3.3	Graph learning	62
4.4	Experiments	63
4.4.1	Experimental procedures	63
4.4.2	Comparison studies	64
4.4.3	Explanatory analysis	66
4.5	Results	67
4.5.1	Comparison studies	67
4.5.2	Explanatory analysis	70

4.5.3	Ablation studies	74
4.6	Discussion	78
5	DreaMR: Diffusion-driven Counterfactual Explanation for Functional MRI	82
5.1	Introduction	83
5.2	Related Work	87
5.2.1	Explanation of fMRI Models	87
5.2.2	Counterfactual Generation	88
5.3	Theory	89
5.3.1	Conventional Diffusion Priors	89
5.3.2	DreaMR	90
5.4	Methods	96
5.4.1	Experimental Procedures	96
5.4.2	Competing Methods	98
5.4.3	Performance Evaluation	101
5.5	Results	102
5.5.1	Fidelity of Counterfactual fMRI Samples	102
5.5.2	Explanations based on Counterfactual Samples	107
5.5.3	Computational Efficiency	110

5.5.4 Ablation Studies	111
----------------------------------	-----

5.6 Discussion	113
--------------------------	-----

6 Conclusion	118
---------------------	------------



List of Figures

- 3.1 (a) Overview of BolT. First, ROI-level BOLD responses are extracted from four-dimensional fMRI data. These responses are then projected by a learnable linear layer to obtain T BOLD tokens. Each BOLD token encodes ROI responses across the brain recorded at a specific time instant as an N -dimensional vector. A cascade of transformer blocks processes BOLD tokens across a collection of F temporally-overlapping windows within the time series. For each time window, a separate learnable classification (CLS) token is employed within the transformer blocks. The CLS tokens input to the first block are initialized as tied vectors across F windows, but they become window-specific following encoding through transformer blocks. The blocks compute latent representations of BOLD and CLS tokens; yet only the CLS tokens are used for the classification task at the output layer. (b) Inner architecture of the transformer block. Unlike vanilla transformers, BolT is equipped with a novel fused window multi-head self-attention (FW-MSA) layer to efficiently capture both local and global context within the fMRI time series. 21

- 3.2 (a) Schematic of the fused window multi-head self-attention (FW-MSA) module. Input BOLD tokens are separated into an overlapping set of time windows of size W and stride s . A CLS token is assigned to each window. A given window possesses a bilateral fringe region of size L to permit interactions with neighboring windows. Within-window interactions are captured via attention among base tokens, whereas cross-window interactions are captured via attention between base and fringe tokens. For each BOLD token, attention-derived latent representations are then fused across the separate time windows in which it appears. (b) Attention calculations within a window. Query to the FW-MSA is a BOLD token from the window base, whereas key and value are BOLD tokens from the broader receptive field including the fringe region. 23
- 3.3 Importance weights for individual cognitive tasks in HCP-Task averaged across subjects (see colorbar). (a) Results for emotion, relational, gambling, language, social and motor tasks. Subtask annotations are shown to outline task structure during fMRI scans. (b) Input fMRI time series and importance weights for the working memory task. Relevant ROIs and time points are emphasized by masking the input time series with importance weights. 36
- 3.4 Accuracy of logistic-regression models receiving as input a subset of five important tokens, while the total importance of the subset is systematically varied. Tokens were ordered according to their importance weights, different subsets of five consecutive tokens in the ordered list were selected. Results are shown for varying subset index for (a) HCP-Task, (b) HCP-Rest, (c) ABIDE-I. Subset index refers to the offset within the ordered list for the selected subset. Importance score was taken as the average importance weight of tokens in a given subset normalized by the minimum importance weight within the time series. As such, total importance of the subset and detection accuracy show a general decrease with increasing subset index. 41

3.5 Landmark time points (i.e. BOLD tokens) selected by BolT were used to identify brain regions critical for gender detection in HCP-Rest. A collection of K=5 tokens were retrieved from each subject, characterizing responses across R ROIs. Next, a logistic regression model was trained to map the tokens in landmark time points onto the associated output class. Model weights reflect each ROI’s contribution to the classification decision. For each class, the top 2 percent of most influential ROIs were visualized (i.e. female in red color, male in blue color). 46

3.6 Critical ROIs for cognitive task detection in HCP-Task. For each task, the top 2 percent of most influential ROIs were visualized. Elevated BOLD responses in highlighted ROIs imply the presence of the associated task. 48

3.7 Critical ROIs for ASD detection in ABIDE-I. The top 2 percent of most influential ROIs were visualized for ASD patients and healthy controls. Elevated BOLD responses in red ROIs imply the presence of the ASD condition, whereas elevated responses in blue ROIs imply a healthy condition. 49

- 4.1 Overview of GraphCorr. A) GraphCorr utilizes two modules in parallel to extract dynamic, lagged features of inter-regional correlations across the brain. The node embedder module receives as input time-windowed BOLD signals, and uses a transformer encoder to compute node embeddings of dynamic FC features $\mathbf{EMB} \in \mathbb{R}^{R \times D \times W}$ (R : number of ROIs, D : embedding dimensionality, W : number of windows). The lag filter module also receives as input time-windowed BOLD signals, albeit it computes lag activations due to cross-correlation between node pairs across a range of time delays $\mathbf{LAG} \in \mathbb{R}^{R \times R \times W \times k}$ (k : number of lag filters). B) To consolidate the extracted feature sets on a graph, node embeddings are taken as node features and lag activations are taken as edge weights. A message passing algorithm is then run on the graph to produce enhanced FC features in an output feature matrix, $\mathbf{OUT} \in \mathbb{R}^{R \times (Dk + D)}$ 57
- 4.2 The node embedder module. The module input is a tensor of time-windowed dynamic FC features $\mathbf{dFC} \in \mathbb{R}^{R \times R \times W}$, with R denoting the number of ROIs, W denoting the number of time windows. The input tensor is processed via a transformer encoder with multi-head self-attention (MHSA), layer normalization (LN), and multi-layer perceptron (MLP) layers. Processing is performed for each time window separately. The module output is a node embedding tensor $\mathbf{EMB} \in \mathbb{R}^{R \times D \times W}$ where D is embedding dimensionality. 60
- 4.3 The lag filter module. The module input is a tensor of time-windowed BOLD signals $\mathbf{F} \in \mathbb{R}^{R \times \Delta T \times W}$, with ΔT denoting the duration of each window. Delayed connectivity features of pairs of ROIs are computed as the cross-correlation of their BOLD signals, with delays $\tau \in \{-m, -m + 1, \dots, m - 1, m\}$. Afterwards, the cross-correlation tensor $\rho \in \mathbb{R}^{R \times R \times W \times (2m+1)}$ is linearly transformed with learnable lag filters $M_{LF} \in \mathbb{R}^{(2m+1) \times k}$ where k denotes the number of filters. The module output is the lag activation tensor $\mathbf{LAG} \in \mathbb{R}^{R \times R \times W \times k}$ 61

4.4 Explanatory analysis for gender detection classifiers. Saliency maps are shown for vanilla, dynamic FC- and GraphCorr-augmented SAGE models on the HCP-Rest dataset, along with reference importance maps derived via a meta-analysis of neuroimaging literature. Important ROIs are marked with colored spheres. In saliency maps, sphere size indicates the relative saliency score for each ROI. In importance maps, sphere size indicates the relative importance score. Color-coding indicates membership to functional networks: (Vis = Visual network, Attn = attention network, SMN = sensorimotor network, Cont = control network, DMN = default mode network.) 71

4.5 Explanatory analysis for cognitive-task detection classifiers. Saliency maps for the motor task are shown based on vanilla, dynamic FC- and GraphCorr-augmented SAGE models on the HCP-Task dataset, along with reference importance maps derived via a meta-analysis of neuroimaging literature. 72

4.6 Analysis of temporal sensitivity for GraphCorr on HCP-Task with the Schaefer atlas. A) As reference for time-varying FC patterns in fMRI scans, conventional dynamic FC (dFC) features were computed via time-windowed correlation measures on BOLD signals. Principal components analysis (PCA) was performed on the dFC features across time windows to obtain a basis for time-varying FC patterns. The first 6 PCs that explain 95% of the variance were selected. B) Static FC (sFC) features and GraphCorr-based FC features were extracted. Since GraphCorr outputs features in a latent embedding space, FC features were derived via correlation measures between the embedding vectors for pairs of ROIs. C) GraphCorr-based FC and sFC features were projected onto the PCs. In general, GraphCorr yields higher projections onto the PCs than sFC. 77

- 5.1 DreaMR is a counterfactual explanation method for deep fMRI classifiers. **a)** To capture the distribution of fMRI data, DreaMR trains a class-agnostic FMD prior that splits the diffusion process into F uniform fractions with dedicated networks $\mathbf{D}_\theta^{[f]}$, as described in Eq. 5.6. Following training, the FMD prior is subjected to multi-phase distillation to attain $\mathbf{D}_{\theta_P}^{[f]}$ that allow fast sampling, as described in Eq. 5.7. **b)** DreaMR is devised to explain the decisions of a deep classifier that predicts cognition-related variables from a subject’s fMRI scan. **c)** Given an input fMRI sample $x_0 \sim p(x)$ mapped onto cognitive state y_0 by the classifier, DreaMR first samples a noise-added version $\bar{x}_{\Delta T}$ via forward diffusion and then generates $\bar{x}_0 \sim p(x)$ with minimal alterations from x_0 via reverse diffusion. Generation is guided with the conditional score of the classifier to flip the decision to $\bar{y}_0 \neq y_0$. Starting at fraction $f_c = \lceil \Delta T(F/T) \rceil$, classifier guidance computed from a denoised estimate of the counterfactual sample ($\tilde{\bar{x}}$) is injected at each diffusion step (orange boxes), as described in Alg. 2. 85

- 5.2 DreaMR leverages a novel FMD prior for efficient sample generation without compromising sample quality, and to enable effective intermittent control over the generation process. The FMD prior splits the overall diffusion process into F fractions, where the f th fraction covers T/F consecutive steps from $t_s(f) = T(f)/F$ to $t_e(f) = T(f - 1)/F + 1$. Since the characteristics of the denoising task can vary notably across fractions, a dedicated denoising network $\mathbf{D}_{\theta_P}^{[f]}$ is employed in each fraction to improve sample fidelity. To enhance sensitivity to long-range temporal context in fMRI scans, the network is built on an efficient transformer architecture that uses fused window attention mechanisms [1]. To improve sampling efficiency, multi-phase distillation is performed over P phases in each fraction (Eq. 5.7). In the p th distillation phase, the diffusion step size is doubled to shorten sampling time by a factor of 2. At the end of multi-phase distillation, the number of sampling steps that must be executed for a given fraction is reduced from T/F to $T/(2^P F)$. Note that the novel combination of diffusion fractions and multi-phase distillation in FMD serves to mitigate losses in sample quality typically encountered with distillation of conventional diffusion priors. 91

5.3 Spatiotemporal explanation maps from competing methods for a representative fMRI scan in HCP-Task based on the Schaefer atlas. The original input fMRI sample for the motor task (left) and the global average of fMRI samples across subjects for the motor task (right) are also shown. For attribution methods, explanation maps were taken as the gradient of the classifier loss function with respect to input features. For perturbation methods, explanation maps were obtained by masking out local patches in fMRI samples. For counterfactual methods, counterfactuals were generated separately to flip the class label from the motor onto each of six remaining cognitive tasks, and explanation maps were taken as the average difference between the original and counterfactual samples. See colorbars for fMRI samples (Signal) and explanation maps (Importance). 98

5.4 Cortical explanation maps produced by DreaMR for female (top panel) and male (bottom panel) sexes on the HCP-Rest dataset. To obtain the cortical maps, counterfactual samples were generated for each original fMRI sample to flip the respective decision of a deep fMRI classifier for sex detection. FC features of original and counterfactual samples were separately derived, and the differences between the two sets of samples were averaged across subjects. Important FC features were determined by selecting the features showing the top 5% of differences. Each important region-of-interest (ROI) in the brain is marked with a dot on the anatomical template, and connectivity between ROIs is shown with a bar. Dot size denotes the importance of the ROI, and bar thickness denotes the importance of the connection. ROIs are colored according to the functional network they belong to (see legend). Lateral and superior brain views are displayed. LH: left hemisphere, RH: right hemisphere; SomMot: somatomotor; DorsAttn: dorsal attention; SalventAttn: Salience/ventral attention. 106

List of Tables

3.1	Performance of BolT variants ablated of essential design elements. Ablated elements were learnable <i>CLS</i> tokens (<i>CLS</i>), split time windows (Windowing), token fusion (Fusion), cross-attention between the base and fringe tokens (Cross Attn.), and cross-window regularization (CWR). When windowing is enabled, the annotation (G) denotes utilization of a global <i>CLS</i> token shared across windows, whereas (L) denotes utilization of local <i>CLS</i> tokens in each window. Results are shown based on both the Schaefer and AAL atlases. Accuracy, recall, precision and AUC metrics are reported as mean(std) across test folds. Bold-face indicates the top-performing model.	31
3.2	Performance of BolT under varying window sizes W . Results are shown based on the Schaefer and AAL atlases. Accuracy, recall, precision and AUC metrics are reported as mean $\pm$ std across test folds. Bold-face indicates the top-performing model.	32
3.3	Performance of logistic-regression models given as input most important tokens determined via the explanatory technique. Results are also given for logistic-regression models based on random tokens. Boldface indicates the top-performing model in each classification task.	37

3.4	Performance of competing methods and BolT using Schaefer atlas for gender detection on the HCP-Rest dataset, task detection on the HCP-Task dataset and disease detection on the ABIDE-I dataset. Metrics are reported as mean $\pm$ std across test folds. Boldface indicates the top-performing model in terms of each metric in individual classification tasks.	42
3.5	Performance of competing methods and BolT using AAL atlas for gender detection on the HCP-Rest dataset, task detection on the HCP-Task dataset and disease detection on the ABIDE-I dataset. Metrics are reported as mean $\pm$ std across test folds. Boldface indicates the top-performing model in terms of each metric in individual classification tasks.	43
4.1	Performance of baseline models for gender detection on HCP-Rest and ID1000, and for cognitive task classification on HCP-Task based on the Schaefer atlas. Accuracy and ROC are listed as mean $\pm$ std across test folds for non-augmented (Vanilla), dynamic FC augmented (Dyn. FC), GRU augmented (GRU), and GraphCorr augmented (GraphCorr) variants. Boldface indicates the top performing plug-in for each model.	69
4.2	Performance of baseline models for gender detection on HCP-Rest and ID1000, and for cognitive task classification on HCP-Task based on the AAL atlas.	70
4.3	Performance for GraphCorr variants ablated of individual design elements in gender detection. Results listed for the augmented SAGE model on the HCP-Rest dataset with the Schaefer atlas. . .	75
4.4	Performance for GraphCorr variants that use $\Delta T = [25\ 1000]$ while $s = 30$, and $s = [25\ 50]$ while $\Delta T = 50$ for gender detection. Results listed for the augmented SAGE model on HCP-Rest with the Schaefer atlas.	76

5.5	Fidelity of spatiotemporal and FC features of counterfactual samples for the HCP-Task dataset. Results based on an aggregate atlas combining ROIs from the MMP and Talairach atlases. . . .	101
5.6	Detection performance of a linear classifier trained using counterfactual samples was measured via accuracy (Acc.) and F1 metrics. Results listed for HCP-Rest, HCP-Task and ID1000 based on the Schaefer atlas.	105
5.7	Detection performance of a linear classifier trained using counterfactual samples for HCP-Task based on the aggregate atlas. Results listed separately for cortical and non-cortical ROIs.	105
5.8	Inference times (Inf., msec) and memory load (Mem., gigabytes) per generation of a counterfactual fMRI sample.	110
5.9	Fidelity of spatiotemporal features of counterfactual samples generated by DreaMR variants formed without a transformer, without Langevin sampling, without fractional diffusion, without multi-phase distillation, and with guidance from classifiers trained on noisy fMRI samples.	111
5.10	Fidelity of spatiotemporal features of counterfactual samples for DreaMR variants formed by implementing the FMD prior with varying number of fractions (F) and number of distillation phases (P).	111

Chapter 1

Introduction

1.1 Deep Learning Applications in Functional MRI Analysis

Functional Magnetic Resonance Imaging (fMRI) has become an indispensable tool for non-invasively studying brain function and its relationship with cognitive processes. By measuring blood-oxygen-level-dependent (BOLD) signals that reflect changes in neural activity, fMRI allows researchers to investigate the complex dynamics of the human brain across various conditions, such as during specific cognitive tasks or at rest. This capability makes fMRI a crucial modality for both basic neuroscience research and clinical applications.

One of the primary approaches in fMRI research involves examining functional connectivity (FC) between different brain regions, which refers to the temporal correlations of BOLD signals across these regions. FC has been extensively studied to understand how brain regions interact during cognitive tasks and in resting states, offering insights into both normal brain function and various neurological conditions. Traditional machine learning methods have been employed to analyze FC features, often using linear models like logistic regression or support vector machines. However, these approaches may fall short in capturing the intricate,

high-dimensional patterns present in fMRI data.

In recent years, deep learning (DL) models have gained traction in fMRI research due to their ability to uncover complex patterns and relationships in large datasets. These models, ranging from convolutional neural networks (CNNs) to graph neural networks (GNNs) and transformers, have demonstrated superior performance over traditional methods. Yet, despite their successes, challenges remain in effectively capturing the dynamic and hierarchical nature of brain activity and in providing interpretable insights from these sophisticated models.

This thesis builds upon the growing body of work in deep learning-based fMRI analysis, contributing novel methods to enhance both the performance and interpretability of fMRI classifiers. The first part of the thesis introduces BolT, a transformer-based model designed to directly operate on BOLD responses for time-series classification. BolT addresses the limitations of traditional and deep-learning models by incorporating a novel fused window attention mechanism, which captures local and global representations across time scales without significantly increasing computational costs.

The second part presents GraphCorr, a plug-in graphical neural network that enhances the sensitivity of fMRI classifiers to dynamic inter-regional interactions. By capturing both instantaneous and lagged interactions among brain regions, GraphCorr provides a more nuanced understanding of functional connectivity, boosting the performance of baseline classifiers on benchmark tasks.

The third and final part of the thesis introduces DreaMR, a deep-learning model that tackles the interpretation challenge in fMRI analysis. DreaMR leverages counterfactual explanations to highlight the critical brain responses that drive the model’s decisions, thereby enhancing the interpretability of deep fMRI models and fostering trust in their use for cognitive assessments.

Collectively, these contributions advance the state-of-the-art in fMRI analysis, offering powerful new tools for understanding the brain’s functional architecture

and providing a pathway toward more interpretable and clinically relevant applications of deep learning in neuroscience.

1.2 Contributions

This thesis presents a series of innovative methodologies to advance the analysis and interpretation of fMRI data, addressing key limitations in existing models and introducing new techniques for more effective brain activity classification.

First, the thesis introduces BolT, a novel transformer architecture designed to efficiently and sensitively analyze fMRI BOLD responses. Traditional approaches rely heavily on convolutional networks, which, while effective in capturing local correlations, often struggle with global context and long-range dependencies. To overcome these challenges, BolT employs a fused window attention mechanism that progressively enlarges window sizes, thereby capturing hierarchical local-to-global representations across the fMRI data. Additionally, a cross-window regularization technique is proposed to align high-level representations across different time points, enhancing the consistency and robustness of the classification features. To facilitate interpretability, an explanatory technique is introduced that assesses the relevance of individual time points and brain regions in the model’s decisions.

Next, the thesis proposes GraphCorr, a novel plug-in designed to enhance the classification performance of baseline models by leveraging dynamic functional connectivity (FC) features from fMRI data. Unlike traditional methods that rely on static FC features, GraphCorr captures temporal variability in brain region connectivity through a message passing algorithm that fuses dynamic FC features across time windows. This fusion reduces feature dimensionality while preserving critical temporal information, enabling more accurate and efficient classification. Furthermore, GraphCorr incorporates a transformer encoder to enable parallel processing of dynamic FC features, and an explicit lag filter mechanism to capture delayed cross-correlations, thereby enhancing the temporal resolution of the

features used in classification.

Finally, this thesis introduces DreaMR, the first diffusion-driven counterfactual explanation method specifically tailored for multivariate fMRI analysis. Conventional diffusion methods are often limited by their inability to efficiently capture the complex spatiotemporal context of fMRI data. To address this, DreaMR utilizes a fractional multi-phase-distilled diffusion prior that improves inference efficiency without sacrificing the quality of the generated counterfactuals. Moreover, DreaMR employs a transformer architecture with linear complexity to effectively model long-range spatiotemporal dependencies within fMRI scans, offering a significant improvement over traditional approaches. This method enables the generation of high-quality counterfactuals that provide deeper insights into the model’s decision-making process.

Through these contributions, this thesis offers a comprehensive set of tools and methodologies that significantly advance the state-of-the-art in fMRI data analysis, paving the way for more accurate, efficient, and interpretable models in neuroscience.

1.3 Outline

The remainder of this thesis is organized as follows: Chapter 2 provides relevant background on MRI and fMRI. Chapter 3 introduces the fused-window transformers model, and Chapter 4 describes the plug-in graph neural network for boosting temporal sensitivity in fMRI analysis. Chapter 5 presents a diffusion-driven counterfactual explanation method for deep fMRI classifiers. Lastly, Chapter 5 summarizes and concludes the main findings and contributions presented in this thesis.

Chapter 2

Fundamentals of Functional Magnetic Resonance Imaging (fMRI)

2.1 Fundamentals of MRI

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that utilizes the principles of nuclear magnetic resonance (NMR) to produce detailed images of the internal structures of the body. This section introduces the fundamental concepts and equations underlying MRI imaging.

2.1.1 Basic Principles

MRI imaging is based on the interaction of magnetic fields with the hydrogen nuclei (protons) in the body. The key components of MRI include:

- **Magnetic Field:** MRI machines generate a strong, uniform magnetic field (B_0), typically in the range of 1.5 to 3.0 Tesla (T). This magnetic field aligns

the magnetic moments of hydrogen nuclei within the body.

- **Radiofrequency (RF) Pulses:** RF pulses are used to perturb the aligned magnetic moments. When these pulses are applied, the protons are excited from their equilibrium state to a higher energy state.
- **Relaxation:** After the RF pulse is turned off, the protons return to their equilibrium state, releasing energy in the process. This energy is detected by the MRI sensors and used to create images. The two primary relaxation processes are:

- **T1 Relaxation (Longitudinal Relaxation):** The time constant T_1 describes the time it takes for the protons to realign with the magnetic field. The equation for T_1 relaxation is given by:

$$M_z(t) = M_0 \left(1 - e^{-\frac{t}{T_1}}\right) \quad (2.1)$$

where $M_z(t)$ is the longitudinal magnetization at time t , and M_0 is the equilibrium magnetization.

- **T2 Relaxation (Transverse Relaxation):** The time constant T_2 describes the decay of the transverse magnetization due to interactions between protons. The equation for T_2 relaxation is:

$$M_{xy}(t) = M_{xy}(0)e^{-\frac{t}{T_2}} \quad (2.2)$$

where $M_{xy}(t)$ is the transverse magnetization at time t , and $M_{xy}(0)$ is the initial transverse magnetization.

2.1.2 Imaging Process

The imaging process in MRI involves several key steps:

- **Magnetic Field Gradients:** MRI uses gradient coils to create variations in the magnetic field, allowing spatial encoding of the MRI signal. Gradient fields are applied along different axes (x, y, and z) to localize the signals in space.

- **Pulse Sequences:** Different pulse sequences are used to highlight various tissue properties. Common sequences include:
 - **Spin Echo (SE):** Provides high-resolution images with T_1 and T_2 weighting.
 - **Gradient Echo (GE):** Offers faster imaging with different weighting based on the T_2^* effect.
- **Fourier Transform:** The acquired MRI signal is transformed from the time domain to the frequency domain using the Fourier transform. This transformation is crucial for accurately mapping spatial information from the signal, thereby enhancing the resolution and quality of the reconstructed images.

2.1.3 Mathematical Modeling

The signal intensity S measured in MRI is influenced by the relaxation times and the applied magnetic field gradients. The general expression for the MRI signal in the presence of gradients can be written as:

$$S = \int \int \int M(x, y, z) \cdot e^{-i\gamma(G_x x + G_y y + G_z z)t} dx dy dz \quad (2.3)$$

where $M(x, y, z)$ represents the magnetization as a function of spatial coordinates, γ is the gyromagnetic ratio, and G_x, G_y, G_z are the gradient strengths in the respective directions.

2.1.4 Machine Learning in MRI

Machine learning has significantly impacted the field of medical imaging, particularly in MRI. The key applications include:

- **Machine Learning in Medical Imaging:** Classical machine learning techniques have been widely applied in medical imaging for various tasks,

including image analysis, reconstruction, and mathematical modeling [2, 3, 4, 5, 6, 7, 8, 9, 10, 11].

- **Deep Learning in Medical Imaging:** In recent years, deep learning techniques have further revolutionized medical imaging. Convolutional neural networks (CNNs) and transformer models, in particular, have enabled more sophisticated image analysis, including automated disease detection, advanced image enhancement, and accurate segmentation [12, 13, 14, 15]. These models automatically learn hierarchical features from the data, significantly improving performance over traditional machine learning approaches [16, 17, 18, 19, 20, 21].
- **MRI Reconstruction:** In MRI, deep learning methods are used to enhance image reconstruction. These methods aim to improve image quality and reduce acquisition time [22, 23, 24, 25, 26]. Deep learning-based reconstruction algorithms can fill in missing data, reduce noise, and correct for artifacts, resulting in higher quality images even with undersampled data [27, 28, 29, 30, 31, 32, 33].
- **MRI Image Synthesis:** Deep learning is also applied to MRI image synthesis, where it generates realistic images from incomplete or low-quality data. Generative models, such as GANs and diffusion models, are employed to create high-fidelity images with realistic tissue contrasts [34, 35, 36, 37]. These models are particularly useful in creating synthetic images that can be used for training purposes or in scenarios where acquiring certain types of MRI images is challenging [38, 39, 40, 41, 42].

2.2 Fundamentals of fMRI

Functional Magnetic Resonance Imaging (fMRI) is an advanced neuroimaging technique that measures brain activity by detecting changes associated with blood flow. This section covers the basics of fMRI imaging, including the underlying principles, key equations, and the concepts of encoding and decoding models in

fMRI.

2.2.1 Basic Principles

fMRI is primarily based on the Blood Oxygen Level-Dependent (BOLD) contrast, which arises from the differing magnetic properties of oxygenated and deoxygenated hemoglobin. When a region of the brain becomes active, the local blood flow to that region increases, leading to a relative decrease in deoxygenated hemoglobin. This change in deoxygenation is detected by fMRI as an indirect measure of neural activity.

- **BOLD Contrast:** The BOLD signal is sensitive to the differences in the magnetic properties of oxygenated (diamagnetic) and deoxygenated (paramagnetic) hemoglobin. The signal intensity $S(t)$ at time t can be described by:

$$S(t) = S_0 [1 + \Delta S(t)] \quad (2.4)$$

where S_0 is the baseline signal intensity, and $\Delta S(t)$ represents the change in signal due to variations in deoxygenated hemoglobin concentration.

The relationship between the BOLD signal and the concentration of deoxygenated hemoglobin $[dHb]$ can be approximated by:

$$\Delta S(t) \propto -\frac{\Delta[dHb](t)}{[dHb]_0} \quad (2.5)$$

where $[dHb]_0$ is the baseline concentration of deoxygenated hemoglobin.

- **Spatial and Temporal Resolution:** fMRI offers relatively high spatial resolution (on the order of millimeters) but limited temporal resolution due to the slow hemodynamic response. The typical temporal resolution ranges from 1 to 3 seconds, which is dictated by the hemodynamic response function (HRF).
- **Hemodynamic Response Function (HRF):** The HRF describes the time course of the BOLD signal following neural activation. The HRF can

be modeled as a convolution of the neural activity $n(t)$ with a hemodynamic response kernel $h(t)$:

$$y(t) = (n * h)(t) = \int_{-\infty}^{\infty} n(\tau)h(t - \tau) d\tau \quad (2.6)$$

where $y(t)$ is the observed BOLD signal. A common model for $h(t)$ is the double-gamma function:

$$h(t) = A \cdot t^{a_1} e^{-t/b_1} - B \cdot t^{a_2} e^{-t/b_2} \quad (2.7)$$

where A and B are scaling constants, a_1 , a_2 are shape parameters, and b_1 , b_2 are time constants.

2.2.2 Encoding Models of fMRI

Encoding models in fMRI aim to describe how stimuli or cognitive processes are represented in brain activity [43, 44]. These models predict the BOLD signal based on the presented stimuli and are critical for understanding brain function.

- **Voxel-wise Modeling:** In voxel-wise encoding models, each voxel's response is modeled based on the stimulus features [45, 46, 47, 48, 49, 50, 51, 52, 43, 53, 52, 54]. For example, in visual cortex studies, models may predict a voxel's response based on features such as orientation or spatial frequency of a visual stimulus. The General Linear Model (GLM) is often employed to relate stimulus features to voxel responses. The GLM is expressed as:

$$Y = X\beta + \epsilon \quad (2.8)$$

where Y is the observed BOLD signal, X is the design matrix representing the stimuli, β are the model coefficients, and ϵ represents the residuals or noise.

The design matrix X can be constructed by convolving the stimulus time course with the HRF:

$$X(t) = S(t) * h(t) \quad (2.9)$$

where $S(t)$ represents the stimulus time course and $h(t)$ is the HRF.

- **Stimulus-Response Models:** These models aim to encode how specific stimuli (e.g., visual, auditory) evoke responses in different brain regions [55, 55, ?]. They are crucial for understanding how sensory information is processed and represented in the brain. The BOLD response in a voxel can be modeled as:

$$Y(t) = \sum_{i=1}^n \beta_i \cdot S_i(t) + \epsilon(t) \quad (2.10)$$

where β_i are the coefficients for each stimulus feature $S_i(t)$.

2.2.3 Decoding Models of fMRI

Decoding models, on the other hand, are used to predict or infer cognitive states or stimuli based on the observed fMRI data. These models are central to applications like brain-computer interfaces and cognitive state classification.

- **Machine Learning Techniques:** Decoding models often leverage machine learning algorithms to classify or predict the stimuli or cognitive states from fMRI data [56]. A common approach is to train a classifier using labeled fMRI data, where the goal is to predict the class label y (e.g., a specific stimulus or cognitive state) from the BOLD signal X . This can be formulated as:

$$y = f(X; \theta) \quad (2.11)$$

where f is the classifier function (e.g., SVM, logistic regression, or a neural network) and θ represents the model parameters.

The training process typically involves minimizing a loss function $L(y, \hat{y})$ that measures the discrepancy between the true labels y and the predicted labels $\hat{y}$:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^N L(y_i, \hat{y}_i) \quad (2.12)$$

where N is the number of samples, and θ^* represents the optimal parameters.

- **Applications of Decoding Models:** Decoding models have been applied in various domains, such as predicting the visual content a person is seeing or identifying the cognitive task a person is performing based on their brain activity. These models provide insights into how information is represented and processed in the brain, offering potential applications in clinical and neuroengineering fields.



Chapter 3

BolT: Fused Window Transformers for fMRI Time Series Analysis

The content of this chapter reflects the studies described in the following publications:

- **H. A. Bedel**, I. Sivgin, O. Dalmaz, S. U. H. Dar, and T. Çukur, "BolT: Fused window transformers for fMRI time series analysis," *Med. Image Anal.*, vol. 88, p. 102841, 2023
- **H. A. Bedel**, I. Sivgin, O. Dalmaz, S. U. H. Dar, and T. Çukur, "Multivariate Classification of fMRI Time Series with Fused Window Transformers," in *31th annual meeting of International Society for Magnetic Resonance Imaging (ISMRM)*, Toronto, June 2023.

3.1 Introduction

Functional MRI (fMRI) measures blood-oxygen-level-dependent (BOLD) responses that reflect changes in metabolic demand consequent to neural activity [57, 58]. Recording BOLD responses at a unique combination of spatio-temporal resolution and coverage, fMRI provides the means to study complex cognitive processes in the human brain [59, 60, 61, 62]. On the one hand, task-based fMRI enables researchers to associate stimulus or task variables with multi-variate responses across the brain [63, 64, 65]. Regions that are co-activated in the presence of a particular variable are taken to be involved in the cortical representation of that variable [66], and they are considered to be functionally connected [67]. On the other hand, characteristic multi-variate responses are also eminent in the absence of external stimuli or task, when the subject is merely resting [68, 69, 70, 71]. In resting-state fMRI, co-activation patterns are typically used to define networks of brain regions, whose functional connectivity (FC) has been associated with various normal and disease states [72, 73, 74, 75]. Many prior studies have linked behavioral traits and prominent neurological diseases with FC features of BOLD responses [76, 77, 78, 79].

Earlier fMRI studies adopted traditional machine learning (ML) to analyze multi-variate brain responses in order to decode task- or disease-related information. Since these ML methods use relatively compact models, feature extraction is typically employed to reduce dimensionality and factor out nuisance variability [80, 81]. A prominent approach first expresses FC features as the temporal correlations of BOLD responses across separate brain regions, and then uses methods such as support vector machines or logistic regression to classify external variables [82, 83, 84, 85]. Later studies have instead adopted deep learning (DL) given its ability to capture complex patterns in high-dimensional data [86, 87, 88, 89, 90]. Various successful deep models have been proposed in the literature based on convolutional [91], graph [92], or recurrent architectures [93, 94] that process FC features. Yet, common FC features primarily reflect first-order inter-regional interactions, potentially disregarding higher-order interactions evident in recorded

BOLD responses [95, 96]. To more directly assess information in fMRI data, several recent studies have instead built classifiers using recurrent models or vanilla transformer models [97, 98, 99] on BOLD responses. While powerful, these recent architectures can introduce high computational burden when processing long time series, and they do not embody explicit mechanisms to capture contextual representations of multi-variate data across diverse time scales [100, 101, 102].

Here we propose a novel transformer architecture that directly operates on BOLD responses, BolT, for fMRI time-series classification. To capture local representations, BolT splits the time series into temporally-overlapping windows and employs a cascade of transformer blocks to encode window-specific representations of BOLD tokens (i.e., linear projections of responses measured across the brain at specific time points). To enhance expressiveness across broad time scales without elevating computational costs, BolT leverages a novel fused window attention mechanism that utilizes cross attention and token fusion among overlapping windows. While cross attention enables interactions between base BOLD tokens in a given window and fringe tokens in neighboring windows prior to encoding, token fusion enables integration of encoded representations across neighboring windows. To hierarchically transition from local to global representations, the extent of window overlap in transformer blocks is progressively increased across the cascade. BolT improves task performance by utilizing classification (*CLS*) tokens to capture task-oriented high-level features. Window-specific *CLS* tokens are introduced to maintain local sensitivity and compatibility with the hierarchical model structure. Meanwhile, task-relevant information exchange is promoted by a novel cross-window regularization that aligns these *CLS* tokens across windows. At the end of the cascade, the encoded *CLS* tokens are averaged across windows and a linear projection layer is used for classification.

Comprehensive demonstrations are reported for classification tasks on public datasets: gender detection from resting-state fMRI scans and cognitive task detection from task-based fMRI scans in the Human Connectome Project (HCP) dataset [103], and disease detection from resting-state fMRI scans in the Autism

Brain Imaging Data Exchange (ABIDE) dataset [104]. BolT achieves higher classification performance than prior traditional and deep-learning methods, including convolutional, graph, recurrent and transformer baselines. Ablation studies are presented that demonstrate the significant contribution of individual design elements to model performance, including learnable *CLS* tokens, split time windows, token fusion, cross attention, and cross-window regularization. To interpret the representational information captured by BolT, we devise an explanatory technique on the fused window attention operators. The proposed technique extracts gradient-weighted attention maps across the cascade to construct an importance map for BOLD tokens, and thereby identify landmark time points. A logistic regression analysis on the landmark points is then performed to identify brain regions that contribute most significantly to the model’s decision. Explanatory analyses reveal task timings and relevant brain regions that corroborate established neuroscientific findings in the literature. Code for implementing BolT is publicly available at <https://github.com/icon-lab/BolT>.

3.2 Related Work

3.2.1 Traditional methods

Whole-brain fMRI data carry densely overlaid patterns of multi-variate responses, which can be difficult to isolate via uni-variate analysis [105, 106]. This has sparked interest in adoption of ML for multi-variate fMRI analysis [107, 108]. Earlier studies in this domain used traditional classifiers such as support vectors machines [109, 110, 111]. Because high-dimensional data are paired with models of limited complexity, feature selection is key to improving sensitivity in traditional models [112, 113, 114]. Accordingly, many traditional models are built on FC features derived from response correlations among brain regions-of-interest (ROIs), as these features are commonly considered to capture discriminative information about cognitive state [115, 116, 117].

3.2.2 Deep learning methods on FC features

In recent years, DL models have been adopted to elevate sensitivity in fMRI analysis. Some studies have used multi-layer perceptron (MLP) or convolutional neural network (CNN) models to extract high-level features of fMRI data [118, 119, 120] and then to classify external variables [121, 122, 123]. More commonly, classification models have been built based on FC features among brain ROIs for improved performance [124, 91, 125]. Given the brain’s intrinsic structure, graph neural networks (GNN) have gained traction wherein individual ROIs denote nodes and FC features among ROIs determine edge weights [126, 127]. To capture temporal variability in dynamic FC features, recurrent or transformer architectures have also been integrated to process the GNN outputs [128]. However, GNN-based models might suffer from over smoothing [129] or squashing [130] that can lower sensitivity to long-range dependencies. Furthermore, while methods that receive FC features as input can improve learning efficiency by mitigating nuisance variability, FC features typically reflect first-order interactions among ROIs, neglecting potential non-linear effects [131].

3.2.3 Deep learning methods on BOLD responses

Recurrent networks: An alternative approach to building models on pre-extracted FC features is to directly analyze BOLD responses in fMRI time series. Given the high degree of temporal correlation in BOLD responses, recurrent neural networks (RNNs) have been proposed to sequentially process fMRI data across time given CNN-based or ROI-extracted spatial representations [132, 97, 133]. Previously reported recurrent architectures in the fMRI literature include vanilla long short-term memory (LSTM) models [97], and hybrid convolutional LSTM models [132, 133]. While recurrent architectures are powerful in time series analysis, sequential processing introduces difficulties in model training on long time series due to vanishing gradients, and hence they may show suboptimal sensitivity to long-range interactions [134].

Vanilla transformers: Transformer architectures based on self-attention mechanisms have recently been introduced to address limitations of recurrent networks [135]. Given a sequence of tokens, self-attention operators filter their inputs based on inter-token similarity to integrate long-range contextual information. A feed-forward network block, typically selected as an MLP, then encodes latent representations of the contextualized tokens. Several recent studies have employed vanilla transformers that process the entire fMRI time series as a single sequence [98, 136]. Vanilla transformers can manifest relatively limited sensitivity to local representations while emphasizing long-range temporal interactions. Moreover, they introduce quadratic computational complexity with respect to sequence length as self-attention requires similarity assessment between all pairs of tokens.

Efficient transformers: To sensitively analyze fMRI data across diverse time scales while mitigating computational burden, here we introduce a novel transformer architecture, BolT, based on a fused window self-attention mechanism (FW-MSA, Section 3.3.1). Unlike previous methods that receive as input pre-extracted FC features [137, 138, 139, 126, 128], BolT performs learning on BOLD responses to improve sensitivity. Unlike vanilla transformers that process the time series as a single sequence to focus on global temporal representations [98, 136], BolT improves efficiency by splitting the time series into overlapping windows, and employs a cascaded transformer encoder that hierarchically extracts local-to-global representations.

Several recent studies have devised efficient transformer models with partially similar aims to our proposed approach. A computer vision study has introduced SwinT that restricts self-attention computations to non-overlapping local windows in a given sequence, and performs half-sequence-length shifts in the window position across alternating transformer blocks [140]. SwinT implicitly captures cross-window interactions via the window shifts, and it does not utilize high-level *CLS* tokens. Instead, BolT explicitly captures cross-window interactions by using overlapping windows along with cross attention and token fusion between neighboring windows, and it utilizes dedicated *CLS* tokens to improve classification.

A natural language processing study has proposed Longformer that restricts self-attention to a moving local window centered on each token for encoding the token in a single context, and it uses a global *CLS* token across the sequence that can degrade sensitivity to local representations [141]. In contrast, BolT encodes each token appearing in multiple overlapping windows in multiple contexts and then fuses these encodings, and it uses window-specific *CLS* tokens that are aligned with cross-window regularization for enhanced sensitivity.

Efficient transformers have also been adopted in medical image analysis tasks. A hybrid CNN-transformer model, IFT-Net, has been proposed that reduces dimensionality of input data with a convolutional module prior to the transformer [142]. While this approach reduces the sequence length for self-attention computations, it can compromise temporal resolution and sensitivity to local representations in time series analysis. Instead, BolT maintains local sensitivity by preserving temporal dimensionality across the cascade. Another study has introduced HATNet as a hybrid CNN-transformer model where the transformer module sequentially computes intra- and inter-window attention on non-overlapping windows [143]. This sequential approach can be suboptimal since inter-window computations are performed on self-attention outputs that ignore cross-window interactions. In contrast, BolT simultaneously computes self- and cross-attention in overlapping windows. A recent fMRI study has introduced a cascaded transformer, TFF, that splits the time series into separate windows to focus on local temporal representations [99]. TFF processes tokens in separate windows independently across the cascade and naively averages the encoded representations over windows, reducing sensitivity to long-range context. Instead, in each stage of the cascade, BolT uses FW-MSA modules to capture interactions that extend over broad time scales via learning-based fusion of information flow across windows.

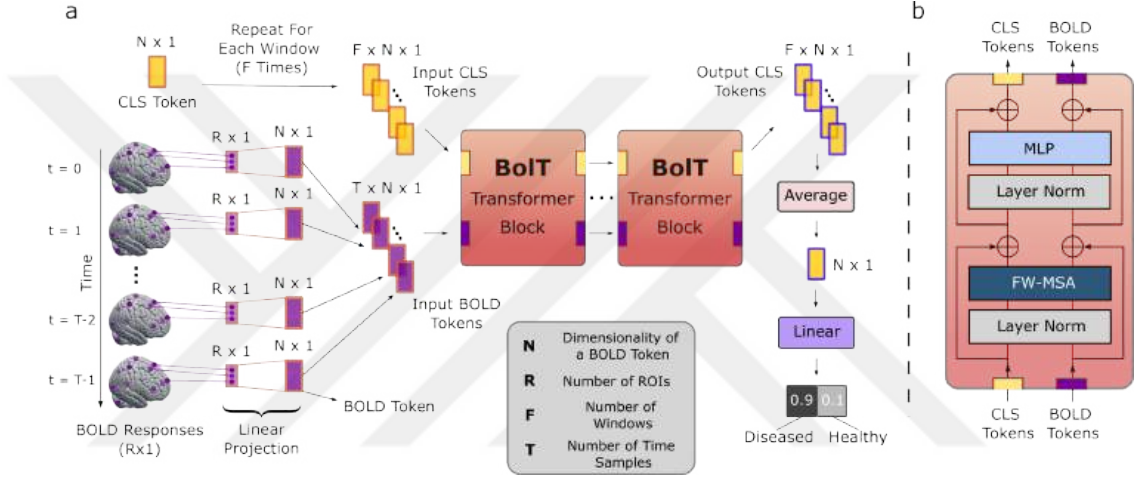


Figure 3.1: (a) Overview of BolT. First, ROI-level BOLD responses are extracted from four-dimensional fMRI data. These responses are then projected by a learnable linear layer to obtain T BOLD tokens. Each BOLD token encodes ROI responses across the brain recorded at a specific time instant as an N -dimensional vector. A cascade of transformer blocks processes BOLD tokens across a collection of F temporally-overlapping windows within the time series. For each time window, a separate learnable classification (CLS) token is employed within the transformer blocks. The CLS tokens input to the first block are initialized as tied vectors across F windows, but they become window-specific following encoding through transformer blocks. The blocks compute latent representations of BOLD and CLS tokens; yet only the CLS tokens are used for the classification task at the output layer. (b) Inner architecture of the transformer block. Unlike vanilla transformers, BolT is equipped with a novel fused window multi-head self-attention (FW-MSA) layer to efficiently capture both local and global context within the fMRI time series.

3.3 Theory

For multi-variate analysis of four-dimensional (4D) fMRI data recorded in a subject, regional BOLD responses are first extracted using an external atlas parcelating the brain into R ROIs. The time series for a given ROI is taken as the average response across voxels within the ROI, z-scored to zero mean and unit variance. Our model learns to map these regional BOLD responses $x \in \mathbb{R}^{T \times R}$ (where T is the number of time samples in the fMRI scan) onto class labels y (e.g. subject gender, cognitive task) depending on the task. Note that transformers expect a sequence of tokens as input. Here, we refer to a learnable linear projection of BOLD responses measured at a particular time index as a BOLD token, i.e., $b^{(t)} = f_b(x^{(t)}) \in \mathbb{R}^N$, where t is the time index, f_b is the linear projection, N is the encoding dimensionality. The collection of BOLD tokens across the fMRI scan is then $b = (b^{(0)}, \dots, b^{(T-1)}) \in \mathbb{R}^{T \times N}$. Latent representations of BOLD tokens are computed by a cascade of transformer blocks in BolT (Figure 3.1). The learned latent features h_f are then linearly projected onto individual class probabilities. To capture both local and global representations of BOLD tokens, the transformer blocks split the fMRI time series into F overlapping windows, and employ a novel fused window self-attention operator to assess interactions between base tokens in a given time window and fringe tokens in neighboring windows. In this section, we describe the architectural details of the proposed model and introduce an explanatory technique devised for BolT.

3.3.1 BolT

BolT embodies a cascade of transformer blocks and a final linear layer to perform classification. Unlike vanilla transformers [135], BolT comprises a novel FW-MSA module to enhance sensitivity to the diverse time scales of dynamic interactions in the brain, while maintaining linear scalability with the duration of fMRI scans (Figure 3.2). A regular MSA layer uses global attention across tokens resulting in quadratic complexity. In contrast, FW-MSA computes local attention within compact time windows extracted from the fMRI scan. Temporal

windowing restricts token-to-token interactions to a focal neighborhood surrounding each window. The resultant local precision serves to improve the capture of subtle changes in brain activation dynamics [144]. To capture a window-level latent representation, a *CLS* token is also employed for use in downstream detection tasks [145]. Input *CLS* tokens are initialized as tied vectors across separate time windows. The *CLS* token for each window is concatenated to the query, value and key tokens in FW-MSA. The final layer uses output *CLS* tokens to linearly map their aggregate features onto class logits.

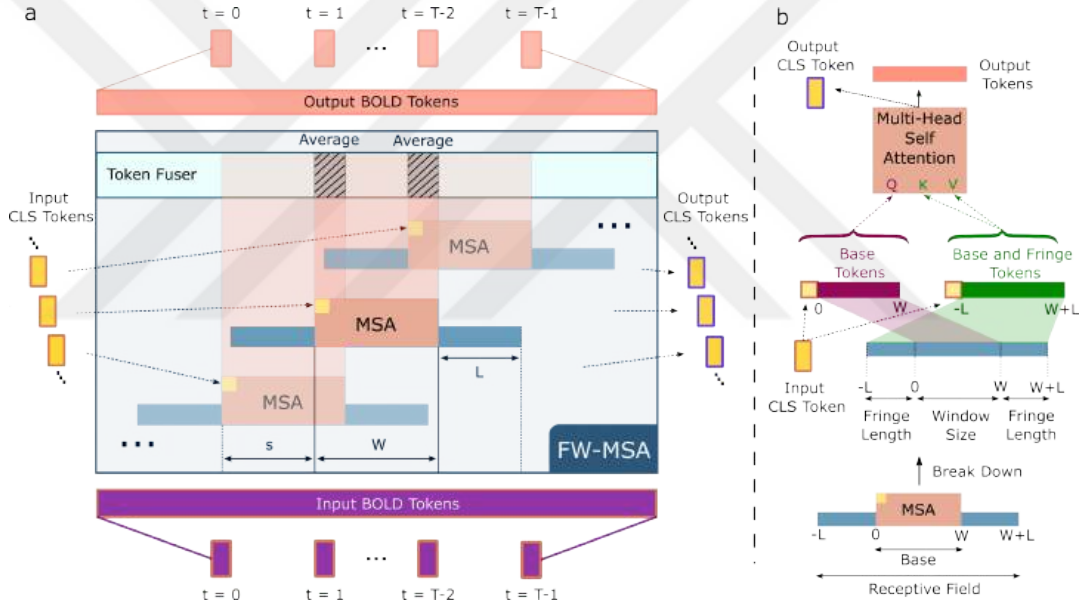


Figure 3.2: (a) Schematic of the fused window multi-head self-attention (FW-MSA) module. Input BOLD tokens are separated into an overlapping set of time windows of size W and stride s . A CLS token is assigned to each window. A given window possesses a bilateral fringe region of size L to permit interactions with neighboring windows. Within-window interactions are captured via attention among base tokens, whereas cross-window interactions are captured via attention between base and fringe tokens. For each BOLD token, attention-derived latent representations are then fused across the separate time windows in which it appears. (b) Attention calculations within a window. Query to the FW-MSA is a BOLD token from the window base, whereas key and value are BOLD tokens from the broader receptive field including the fringe region.

Fused window attention: FW-MSA enables cross-window interactions by attention calculation between base tokens in a given window and fringe tokens in neighboring windows. To do this, FW-MSA first splits the entire collection

of BOLD tokens $b \in \mathbb{R}^{T \times N}$ into $F = (T - W)/s + 1$ windows of size W and stride s . The receptive field of a given window contains W base tokens centrally, and L fringe tokens on either side of the base. While processing the i -th window, FW-MSA receives as input a collection of CLS ($CLS_i \in \mathbb{R}^N$) and BOLD ($b_i \in \mathbb{R}^{(W+2L) \times N}$) tokens. Let $Q_i \in \mathbb{R}^{(1+W) \times N}$ denote queries for base tokens, and $K_i \in \mathbb{R}^{(1+W+2L) \times N}$ and $V_i \in \mathbb{R}^{(1+W+2L) \times N}$ denote keys and values for the union of base and fringe tokens. Assuming f_q, f_k and f_v are learnable linear projections, the query, key and value for the i -th window are:

$$\begin{aligned} Q_i &= f_q(\{CLS_i, b^{(i \times s)}, \dots, b^{(i \times s + W - 1)}\}), \\ K_i &= f_k(\{CLS_i, b^{(i \times s - L)}, \dots, b^{(i \times s + W + L - 1)}\}), \\ V_i &= f_v(\{CLS_i, b^{(i \times s - L)}, \dots, b^{(i \times s + W + L - 1)}\}). \end{aligned} \quad (3.1)$$

To leverage information in the temporal ordering of BOLD tokens, we incorporate a relative position bias in attention calculations [140, 146]:

$$\text{Attention}(Q_i, K_i, V_i) = \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}} + B\right) V_i, \quad (3.2)$$

where $B \in \mathbb{R}^{(1+W) \times (1+W+2L)}$ is a learnable positional bias matrix and d is the feature dimensionality of the attention head. FW-MSA includes multiple attention heads, albeit expressions are given for a single head for simplicity. B expresses the positioning of base and CLS tokens with respect to all tokens in the receptive field including base, fringe and CLS tokens. For BOLD tokens, B parametrizes the potential range $[-W - L + 1, W + L - 1]$ of relative distances among tokens in different positions of the receptive field. For the CLS token, it instead serves to distinguish the CLS token from BOLD tokens.

Token fusion: FW-MSA calculates latent representations of each BOLD token given surrounding local context. Because each token appears in multiple windows, a token fuser is used to aggregate the resultant representations:

$$b^{(i)}[m] = \frac{1}{P} \sum_{p=0}^{P-1} b_p^{(i)}[m - 1], \quad (3.3)$$

where $m \in \{0, 1, \dots, M - 1\}$ is the index of the transformer block, p is the index of the time window among P windows that contain a particular token, $b_p^{(i)}[m -$

1] is the i -th input BOLD token, and $b^{(i)}[m]$ is the fused token. Token fusion facilitates exchange of information across windows, while maintaining a fixed number of BOLD tokens across transformer blocks. Following fusion, all tokens are forwarded to the MLP module.

Cross-window regularization: The first transformer block in BolT receives as input a single CLS token shared across the time windows. The transformer encoders then compute a unique CLS token for each window based on the BOLD tokens within its receptive field. If the latent space of window-level representations captured by the CLS tokens are largely incompatible, model performance in downstream classification tasks might be compromised. Thus, to encourage high-level representations that are consistent across time windows, we introduce a novel cross-window regularization as an additional loss term:

$$L_{CWR} = \frac{1}{NF} \sum_{i=0}^{F-1} \left\| CLS_i[M-1] - \frac{1}{F} \left(\sum_{j=0}^{F-1} CLS_j[M-1] \right) \right\|_2^2, \quad (3.4)$$

where $CLS_i[M-1]$ is the encoded CLS token for the i -th window at the output of the last transformer block (M -th). The regularization term in Eq. 3.4 penalizes the deviation of individual CLS tokens from their mean over windows.

3.3.2 Explanatory technique

We introduce an explanatory technique for BolT that generates importance weights for BOLD tokens to assess their contribution to a given decision. To do this, we first derive gradient-weighted attention maps as inspired by recent computer vision studies [147]. Yet, we propose an adapted procedure for map calculation in FW-MSA to cope with the overlap and fusion operations across time windows (Alg. 1). In the proposed procedure, gradient-weighted attention maps are calculated separately for each transformer block and each time window:

$$\bar{A}_{mi} = E_h((\nabla A_{mi} \odot A_{mi})^+). \quad (3.5)$$

In Eq. 3.5, $A_{mi} \in \mathbb{R}^{(1+W) \times (1+W+2L)}$ is the attention map produced by the FW-MSA layer at the m -th block for the i -th window, and the first row and column

of A_{mi} are reserved for attention values related to the CLS token. The effect of map values onto the model output is characterized via ∇A_{mi} , i.e., the gradient of the loss function with respect to A_{mi} . Meanwhile, E_h denotes the averaging operator across attention heads for aggregation, $\odot$ is the Hadamard product to modulate the attention maps with the gradients, and $^+$ denotes rectification to prevent negative values. Within each transformer block, single-window attention maps are then aggregated to form a global attention map across the entire time series, $\bar{A}_G[m] \in \mathbb{R}^{(F+T) \times (F+T)}$ where F is the number of windows, T is the number of BOLD tokens. During aggregation, projections of fringe tokens that appear in multiple windows are averaged across windows. Assuming that $t' = F + i \times s$ is the starting index of base tokens in the i -th window, $t'' = t' - L + p_l$ and $t''' = t' + W + L - p_r$ respectively denote the starting and ending indices of the fringe tokens, where $p_l = \max(0, i \times s - L) - (i \times s - L)$ and $p_r = i \times s + W + L - \min(T, i \times s + W + L)$ are correction factors to handle windows near the edges of the time series. Receiving as input $\bar{A}_{mi}$, the projection in the i -th window is then expressed as:

$$\bar{A}_G[m](i, i) = \bar{A}_{mi}(0, 0) \quad (3.6)$$

$$\bar{A}_G[m](i, t'' : t''') = \bar{A}_{mi}(0, 1 + p_l : -p_r) \quad (3.7)$$

$$\bar{A}_G[m](t' : t' + W, i) = \bar{A}_{mi}(1 :, 0) \quad (3.8)$$

$$\bar{A}_G[m](t' : t' + W, t'' : t''') = \bar{A}_{mi}(1 :, 1 + p_l : -p_r) \quad (3.9)$$

$$\bar{A}_G[m] = \bar{A}_G[m] \odot A_{norm} \quad (3.10)$$

Eq. (3.6) captures self-attention for CLS_i , Eq. (3.7) captures attention between CLS_i and BOLD tokens within the receptive field of the i -th window, Eq. (3.8) captures attention between base BOLD tokens and the CLS_i token, and Eq. (3.9) captures attention between based and fringe BOLD tokens. Note that $(a : -b + 1)$ selects between the $(a + 1)$ -th element from the start and the b -th element from the end. In Eq. (3.10), $A_{norm} \in \mathbb{R}^{(F+T) \times (F+T)}$ is an occurrence matrix that captures the number of times each token occurs across windows (i.e., all entries for a given token that appears in n windows are set to n), and $\odot$ is Hadamard division used to normalize for repeated token occurrence.

Algorithm 1: Calculation of relevancy map

Input: $\{\{A_{0(0)}, \dots, A_{0(F-1)}\}, \dots, \{A_{M-1(0)}, \dots, A_{M-1(F-1)}\}\}$: Set of F single-window attention maps from FW-MSA modules across M transformer blocks.

Output: $Rel[M]$: Relevancy map.

$Rel[0] = \mathbf{I}_{F+T}$ *Initialize relevancy map*

for $m = 0 : M - 1$ **do**

$\bar{A}_G[m] = \mathbf{0}_{F+T}$ *Initialize global attention map*

for $i = 0 : F - 1$ **do**

$\bar{A}_{mi} \leftarrow E_h((\nabla A_{mi} \odot A_{mi})^+)$ *Weighted attention map*

$t' \leftarrow F + i \times s$

$p_l \leftarrow \max(0, i \times s - L) - (i \times s - L)$

$p_r \leftarrow i \times s + W + L - \min(T, i \times s + W + L)$

$t'' \leftarrow t' - L + p_l$

$t''' \leftarrow t' + W + L - p_r$

$\bar{A}_G[m](i, i) \leftarrow \bar{A}_{mi}(0, 0)$ *CLS to CLS attention*

$\bar{A}_G[m](i, t'' : t''') \leftarrow \bar{A}_{mi}(0, 1 + p_l : -p_r)$ *CLS to BOLD*

$\bar{A}_G[m](t' : t' + W, i) \leftarrow \bar{A}_{mi}(1 :, 0)$ *BOLD to CLS*

$\bar{A}_G[m](t' : t' + W, t'' : t''') \leftarrow \bar{A}_{mi}(1 :, 1 + p_l : -p_r)$ *BOLD to BOLD*

$\bar{A}_G[m] \leftarrow \bar{A}_G[m] \odot A_{norm}$ *Normalize for repeats*

$Rel[m+1] \leftarrow Rel[m] + \bar{A}_G[m]Rel[m]$ *Update rel. map*

return $Rel[M]$

Next, a token-relevance map $Rel[0]$ that represents the influence of each token onto other tokens in the time series is initialized as an identity matrix in $\mathbb{R}^{(F+T) \times (F+T)}$, implying that each token is initially self-relevant. The normalized attention maps are then used to progressively update the token-relevancy map across transformer blocks where $m \in [0, 1, \dots, M - 1]$:

$$Rel[m+1] = Rel[m] + \bar{A}_G[m]Rel[m]. \quad (3.11)$$

Following the calculation of the token-relevancy map at the final FW-MSA module, importance weights for input BOLD tokens $w_{imp} \in \mathbb{R}^T$ are finally derived as:

$$w_{imp} = \frac{1}{F} \sum_{i=0}^{F-1} Rel[M](i, F :). \quad (3.12)$$

Importance weight of a BOLD token for the classification task is taken as the across-window average of relevancy scores between the *CLS* tokens and the given BOLD token.

3.4 Methods

3.4.1 Datasets

Demonstrations were performed on fMRI data from the HCP S1200¹ [103] and ABIDE I releases² [104]. In HCP S1200, resting-state fMRI data (HCP-Rest) were analyzed to predict gender, and task-based fMRI data (HCP-Task) were analyzed to predict cognitive task. In ABIDE I, resting-state fMRI data were analyzed to detect Autism Spectrum Disorder (ASD). Details about datasets are provided below.

HCP-Rest: Preprocessed fMRI data from 1200 subjects released by the WU-Minn HCP consortium were analyzed [148]. For each subject, the first session of resting-state scans was used, and incomplete scans with shorter than 1200 time samples were excluded. HCP-Rest comprised a total of 1093 scans from 594 female and 499 male subjects.

HCP-Task: Preprocessed fMRI data from 1200 subjects released by the WU-Minn HCP consortium were analyzed [148]. The first session of task-based scans was used, where each subject performed seven different tasks in separate runs: emotion, relational, gambling, language, social, motor, and working memory. Incomplete scans were excluded. HCP-Task comprised a total of 7450 scans from 594 female and 501 male subjects.

ABIDE-I: Preprocessed fMRI data released by the Preprocessed Connectomes Project were analyzed [149, 104]. Low-quality scans that did not pass quality checks from all raters were excluded from analysis. ABIDE-I comprised a total of 871 scans from 403 patients with ASD and 468 healthy controls [137].

¹<https://db.humanconnectome.org>

²https://fcon_1000.projects.nitrc.org/indi/abide/

3.4.2 Experimental procedures

Experiments were conducted in PyTorch on an NVIDIA RTX 3090 GPU. Modeling was performed via a nested cross-validation procedure with 10 outer and 1 inner folds. Accordingly, subjects were split into non-overlapping training (80%), validation (10%), and test sets (10%). For fair comparison, all competing methods used identical data splits. For each competing method, hyperparameter selection was performed based on performance in the first validation set and selected parameters that showed near optimal performance across all datasets and atlases were used thereafter. The selected parameters included learning rate $\in (10^{-6}, 10^{-1})$, number of epochs $\in (5, 100)$ and mini-batch size $\in (1, 100)$. Training was performed via the Adam optimizer. BolT was trained to minimize the following loss: $L = L_{CE} + \lambda \cdot L_{CWR}$ where L_{CE} is cross-entropy loss, and $\lambda = 0.1$ is the regularization coefficient for CWR loss set via cross-validation. Mean and standard deviation of model performance were reported across the test sets.

For each subject in the training set, fMRI time series were randomly cropped in the temporal dimension to 600 samples for HCP-Rest, 150 samples for HCP-Task, and 100 samples for ABIDE-I to improve stochasticity and learning efficiency [128]. Functional data were registered to corresponding structural data for each subject, and aligned to the MNI template. ROI definitions were implemented using two public brain atlases: the Schaefer atlas [150] with 400 regions labeled across seven intrinsic connectivity networks, and the AAL atlas [151] with 116 regions.

3.4.3 Implementation details

In this section, the architectural and hyperparameter details of BolT are summarized. BolT was trained for 20 epochs with a batch size of 32 and an initial learning rate of 10^{-4} . The learning rate was increased to 2×10^{-4} in the first 10 epochs and then gradually decreased to 10^{-5} . A linear projection layer matched the dimensionality of input BOLD responses to the hidden dimensionality of the

transformer blocks. A cascade of four blocks was used, each composed of FW-MSA and MLP modules that used layer normalization and skip connections. A hidden dimensionality of 400 and 40 attention heads with 20 dimensions per head were prescribed. A dropout rate of 0.1 was used in both FW-MSA and MLP layers. For FW-MSA modules, given a desired window size W , stride s and fringe length L were set proportionately as follows:

$$s = W\alpha, \quad L = m(W - s)\beta = m(1 - \alpha)W\beta, \quad (3.13)$$

where $m \in \{0, 1, \dots, M - 1\}$ is the block index, $\alpha \in \mathbb{R}^+$ is the stride coefficient (i.e., proportionality constant), and $\beta \in \mathbb{Z}^+$ is the fringe coefficient. Note that the fringe length was progressively grown over transformer blocks as the number of fused tokens increased. Here, cross-validated search for $W \in (10, 200)$, $\alpha \in (0.0, 1.0)$, $\beta \in (0, 3)$ was performed. Hyperparameters were selected as $W = 20$, $\alpha = 0.4$, $\beta = 2$.

Table 3.1: Performance of BolT variants ablated of essential design elements. Ablated elements were learnable *CLS* tokens (*CLS*), split time windows (Windowing), token fusion (Fusion), cross-attention between the base and fringe tokens (Cross Attn.), and cross-window regularization (CWR). When windowing is enabled, the annotation (G) denotes utilization of a global *CLS* token shared across windows, whereas (L) denotes utilization of local *CLS* tokens in each window. Results are shown based on both the Schaefer and AAL atlases. Accuracy, recall, precision and AUC metrics are reported as mean(std) across test folds. Bold-face indicates the top-performing model.

Atlas	<i>CLS</i>	Windowing	Fusion	Cross Attn.	CWR	Acc. (%)	Rec. (%)	Prec. (%)	AUC (%)
Schaefer	✗	✗	✗	✗	✗	86.35 ±3.56	85.76 ±4.94	84.74 ±4.68	94.59 ±1.57
	✓	✗	✗	✗	✗	89.65 ±3.36	88.38 ±5.63	89.15 ±4.78	96.03 ±1.23
	✓(G)	✓	✗	✗	✗	84.99 ±2.37	89.78 ±4.76	79.95 ±2.68	93.63 ±1.58
	✓(L)	✓	✗	✗	✗	89.65 ±1.85	88.77 ±3.27	88.71 ±3.06	96.79 ±0.90
	✓(L)	✓	✓	✗	✗	90.29 ±1.77	89.57 ±2.54	89.34 ±3.23	97.15 ±1.01
	✓(L)	✓	✓	✓	✗	91.03 ±2.12	89.97 ±2.70	90.42 ±3.30	97.09 ±1.10
	✓(L)	✓	✓	✓	✓	91.85 ±3.05	90.58 ±4.97	91.51 ±3.07	97.35 ±1.06
AAL	✗	✗	✗	✗	✗	80.01 ±3.06	78.20 ±5.61	78.22 ±3.88	88.74 ±3.14
	✓	✗	✗	✗	✗	83.12 ±3.58	80.80 ±5.38	82.25 ±5.05	90.88 ±1.91
	✓(G)	✓	✗	✗	✗	78.74 ±4.11	85.00 ±5.07	73.02 ±4.41	88.41 ±3.41
	✓(L)	✓	✗	✗	✗	86.85 ±3.06	87.20 ±4.30	84.78 ±4.65	93.57 ±2.25
	✓(L)	✓	✓	✗	✗	87.04 ±2.48	87.60 ±4.96	84.73 ±3.43	93.80 ±2.13
	✓(L)	✓	✓	✓	✗	87.13 ±3.03	85.80 ±5.39	86.17 ±4.22	94.12 ±1.98
	✓(L)	✓	✓	✓	✓	87.31 ±2.69	86.99 ±4.49	85.65 ±4.01	94.29 ±2.05

Table 3.2: Performance of BolT under varying window sizes W . Results are shown based on the Schaefer and AAL atlases. Accuracy, recall, precision and AUC metrics are reported as mean $\pm$ std across test folds. Bold-face indicates the top-performing model.

Atlas	Window size	Acc. (%)	Rec. (%)	Prec. (%)	AUC (%)
Schaefer	W=10	91.39 $\pm$ 2.67	89.98 $\pm$ 4.28	91.09 $\pm$ 2.81	97.49 $\pm$ 0.78
	W=20	91.85$\pm$3.05	90.58$\pm$4.97	91.51$\pm$3.07	97.35 $\pm$ 1.05
	W=80	91.48 $\pm$ 2.17	90.18 $\pm$ 4.41	91.18 $\pm$ 2.86	97.56$\pm$0.95
	W=200	88.74 $\pm$ 2.73	84.77 $\pm$ 5.92	90.04 $\pm$ 2.34	96.32 $\pm$ 1.44
AAL	W=10	86.58 $\pm$ 2.68	86.20 $\pm$ 4.77	84.88 $\pm$ 3.98	93.83 $\pm$ 2.10
	W=20	87.31$\pm$2.69	86.99$\pm$4.49	85.65$\pm$4.01	94.29$\pm$2.05
	W=80	86.39 $\pm$ 4.42	86.00 $\pm$ 4.56	84.71 $\pm$ 5.88	94.17 $\pm$ 2.30
	W=200	85.57 $\pm$ 4.04	82.80 $\pm$ 5.60	85.40 $\pm$ 5.31	91.98 $\pm$ 3.72

3.4.4 Model complexity

Complexity of models that directly analyze BOLD responses as a temporal sequence depends on the extent of computations performed on the input sequence of BOLD tokens. Assume that the latent dimensionality of tokens is N in a sequence of length T . Recurrent architectures process the tokens in the input sequence serially. As such, a recurrent layer follows a well-known $O(N^2T)$ complexity that scales linearly with the sequence length [97, 133, 125]. A convolution layer also has a linear $O(kN^2T)$ complexity where k denotes the size of the convolution kernel. Meanwhile, vanilla transformers use regular MSA layers that exhaustively compute interactions between all time points in the input sequence, with T queries and T keys of dimensionality N . Thus, vanilla transformers such as BAnD and IFT-Net incur $O(NT^2)$ complexity that scales quadratically with the input sequence length [98, 142, 152].

In contrast, the FW-MSA layer in BolT computes focal interactions between

time points in overlapping time windows. Each window has $W + 1$ queries and $W + 2L + 1$ keys. The complexity within a single window is $O(NW^2 + NWL)$. Given a total of $(T - W)/s$ windows for the entire sequence, the overall complexity is $O(NTW^2/s + NTWL/s)$. Selecting s and L as outlined in Eq.3.13, BolT incurs $O(NTW^{\frac{(1+\beta(m)(1-\alpha))}{\alpha}})$ complexity that linearly scales with sequence length. Other efficient transformers also show a similar linear trend [140, 99, 141, 143]. For instance, $L = 0$ (no cross-window attention) and $s = \frac{W}{2}$ result in a linear $O(NTW)$ complexity in SwinT [140], and $L = 0$ and $s = W$ result in $O(NTW)$ complexity in TFF and HATNet [99].

3.4.5 Competing methods

BolT was demonstrated against several state-of-the-art methods for fMRI classification including recent transformer, graph, convolutional and recurrent network models, along with a traditional classifier. The architecture, loss function and learning rate scheduler for each competing method were adopted from the original proposing papers.

SVM: A traditional model operating on static FC features was considered [137]. An ℓ_2 regularized model with linear kernel was used. FC features were computed via Pearson’s correlation between ROI-level responses. The cross-validated hyperparameter was a regularization weight of $C=1$.

BrainNetCNN: A CNN model operating on static FC features of fMRI data was considered [91]. ROI-level features were processed with a cascade of two edge-to-edge, one edge-to-node, and one node-to-graph convolutional layers followed by three linear layers. FC features were computed via Pearson’s correlation. Cross-validated hyperparameters were 10^{-4} learning rate, 20 epochs, and 16 batch size.

BrainGNN: A GNN operating on static FC features of fMRI data was considered [126]. Taking ROIs as graph nodes, BrainGNN used a cascade of graph convolutional layers to assign nodes to clusters with learned embeddings, and

used pooling layers to aggregate information with element-wise score normalization. FC features were computed via partial correlation [126]. Cross-validated hyperparameters were 10^{-2} learning rate, 50 epochs, and 100 batch size.

STAGIN: A GNN operating on dynamic FC features of fMRI data was considered [128]. STAGIN processed shifted windows across the time series by a four-layer GNN taking ROIs as nodes, deriving edges based on FC features, and calculating node features via a recurrent unit. Features extracted from each graph by a squeeze-excitation readout module were consolidated onto a single latent feature by a transformer. The latent feature was linearly projected to class logits. A window size of 50 was prescribed, FC features were computed via Pearson’s correlation [128]. Cross-validated hyperparameters were 2×10^{-4} learning rate, 40 epochs, 8 batch size.

LSTM: An RNN model operating on BOLD responses was considered [97]. A single-layer LSTM model averaged hidden states across time samples, and performed classification via a sigmoid activation layer. Cross-validated hyperparameters were 10^{-3} learning rate, 30 epochs, 64 batch size.

CNN-LSTM: A convolutional RNN model operating on BOLD responses was considered [133]. CNN-LSTM was adopted to use 1D convolutions. Hidden states were averaged across time samples to perform classification via a sigmoid activation layer. Cross-validated hyperparameters were 10^{-3} learning rate, 50 epochs, and 64 batch size.

GC-LSTM: A graph convolutional RNN model operating on FC features was considered [125]. GC-LSTM was adopted to use the FC features to construct graph adjacency matrices on windowed time series, and it computed recurrent state updates via spectral graph convolution. The window size and stride were set to 120 and 2 sec, respectively [125]. Hidden states were averaged across time samples to perform classification via an MLP followed by a sigmoid activation layer [125]. Cross-validated hyperparameters were a learning rate of 10^{-3} , 100 epochs, and a batch size of 16.

SwinT: A transformer model with a windowed attention mechanism proposed for computer vision tasks was considered [140]. For fair comparison, the number and hidden dimensionality of transformer blocks, number of attention heads, and window size were matched with BolT. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer blocks. Output tokens were averaged across windows and linearly projected onto class logits for classification [141]. Cross-validated hyperparameters were 2×10^{-4} learning rate, 30 epochs, and 32 batch size.

Longformer: A transformer model with a windowed attention mechanism proposed for language tasks was considered [141]. Longformer uses sliding window attention for BOLD tokens, and global attention for the *CLS* token. For fair comparison, the number and hidden dimensionality of transformer blocks, number of attention heads, and window size were matched with BolT. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer blocks. The global *CLS* token output was linearly projected onto class logits for classification [141]. Cross-validated hyperparameters were 2×10^{-4} learning rate, 40 epochs, and 32 batch size.

BAnD: A hybrid CNN-transformer model operating on BOLD responses was considered [98]. The CNN module suffered from suboptimal learning, so for fair comparison BAnD was implemented with ROI-level inputs as in BolT. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer. No windowing was performed on the time series. Cross-validated hyperparameters were 10^{-4} learning rate, 30 epochs, and 32 batch size.

TFF: A hybrid CNN-transformer model operating on BOLD responses was considered [99]. The CNN module was observed to suffer from suboptimal learning, so for fair comparison TFF was implemented with ROI-level inputs as in BolT. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer. The time series was split into windows of size 20 with shifts of 10, which were processed independently until the final layer where they were averaged [99]. Cross-validated hyperparameters were 10^{-5} learning rate, 30 epochs, and 32 batch size.

IFT-Net: A hybrid CNN-transformer model proposed for medical image analysis was considered [142]. IFT-Net was adopted to use 1D convolutional projections in attention layers where ROI-level responses were taken as input channels. ROI-level responses were linearly projected to match the hidden dimensionality of the transformer blocks. The output feature vector was fed to a sigmoid activation layer for classification. Cross-validated hyperparameters were 10^{-4} learning rate, 30 epochs, and 8 batch size.

HATNet: A hybrid CNN-transformer model proposed for medical image analysis was considered [143]. HATNet was adopted to use 1D CNN modules with ROI-level responses taken as input channels. The output of the CNN module was split into windows of size 16 and processed via the transformer module for classification [143]. Cross-validated hyperparameters were 10^{-4} learning rate, 100 epochs, and 32 batch size.

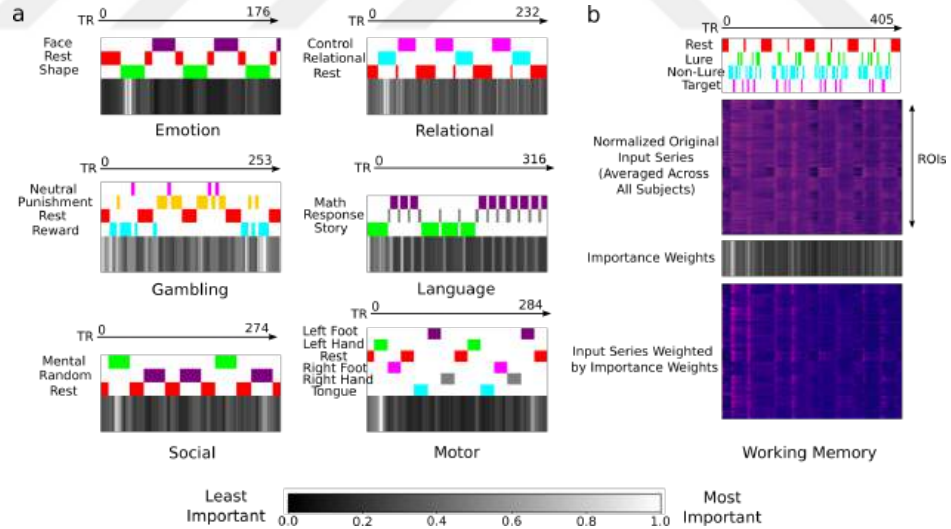


Figure 3.3: Importance weights for individual cognitive tasks in HCP-Task averaged across subjects (see colorbar). (a) Results for emotion, relational, gambling, language, social and motor tasks. Subtask annotations are shown to outline task structure during fMRI scans. (b) Input fMRI time series and importance weights for the working memory task. Relevant ROIs and time points are emphasized by masking the input time series with importance weights.

Table 3.3: Performance of logistic-regression models given as input most important tokens determined via the explanatory technique. Results are also given for logistic-regression models based on random tokens. Boldface indicates the top-performing model in each classification task.

	LR (Important)	LR (Random)
HCP-Rest		
Acc. (%)	77.74±6.82	49.59±2.79
Rec. (%)	71.93±6.43	26.86±3.76
Prec. (%)	77.98±8.47	41.98±4.37
AUC (%)	83.70±5.64	45.70±1.25
HCP-Task		
Acc. (%)	94.83±1.44	15.14±1.18
Rec. (%)	94.81±1.44	15.15±1.18
Prec. (%)	94.89±1.43	15.51±1.47
AUC (%)	99.22±0.33	51.19±0.73
ABIDE-I		
Acc. (%)	60.44±3.55	51.23±4.05
Rec. (%)	46.81±4.80	33.11±5.36
Prec. (%)	59.71±4.84	47.68±6.31
AUC (%)	62.26±4.65	49.33±5.34

3.5 Results

3.5.1 Ablation studies

We conducted a series of ablation studies to evaluate the contribution of individual design elements in BolT. The design elements included learnable and local (i.e., window-specific) *CLS* tokens, split time windows, token fusion, cross attention between base and fringe tokens, and cross-window regularization. Starting

with a vanilla transformer variant, ablated variants were obtained by progressively introducing individual elements. The vanilla variant omitted all elements including *CLS* tokens, so classification was performed by linearly projecting the time average of encoded BOLD tokens at the output of the last transformer block. For all variants, the architecture and hyperparameters for the utilized components were matched with BolT. To assess the contribution of a learnable *CLS* token, a variant was formed by introducing a global *CLS* token into the vanilla variant. Note that local *CLS* tokens are not applicable in this case since windowing was omitted. To assess the contribution of windowing, two variants were formed that used split time windows and either global or local *CLS* tokens, albeit omitted token fusion, cross attention and cross-window regularization. Since cross attention was not used, these variants split the time series into non-overlapping windows by setting stride equal to window size and fringe coefficient to 0 (i.e., $s = W$, $L = 0$). For the global *CLS* variant, attention for the *CLS* token was computed with all BOLD tokens in the time series, whereas attention for a BOLD token was restricted to the window it resided in. For the local *CLS* variant, attention computations for all tokens were restricted to local windows. To assess the contribution of token fusion, a variant was formed that used local *CLS* tokens, splitting into overlapping windows (i.e., $s = W\alpha$ as in BolT) and token fusion, albeit omitted cross attention and cross-window regularization. Cross attention was omitted by setting $L = 0$. To assess the contribution of cross attention, a variant was formed that used local *CLS* tokens, splitting into overlapping windows, token fusion and cross-window attention (i.e., $L = m(1 - \alpha)W\beta$), albeit omitted cross-window regularization. Finally, to assess the contribution of cross-window regularization, a variant using all elements (i.e., BolT) was employed where regularization was achieved via adding the loss term in Eq. 3.4.

Table 3.1 lists performance metrics for all ablated variants. First, we find that introduction of a learnable, global *CLS* token into the vanilla variant consistently improves performance metrics, demonstrating the utility of high-level *CLS* over low-level BOLD tokens in classification tasks. Second, a performance loss is incurred when the global *CLS* token is used along with split time windows, suggesting that the global token does not adequately capture local representations.

In contrast, local *CLS* tokens in combination with windowing yield a performance boost, indicating the importance of using window-specific *CLS* tokens to learn local representations. Third, fusion of tokens with repeated occurrence in overlapping windows yields a further improvement. Note that the BOLD token for a given time point appears in a single window and is encoded in a single context for a non-overlapping split, whereas it appears in multiple windows and is encoded in multiple contexts for an overlapping split. Thus, improvements due to token fusion demonstrate the benefit of encoding representations of time points in diverse contexts. Fourth, enabling cross attention between base tokens in each window and neighboring fringe tokens increases performance, indicating the importance of this cross-attention mechanism for integration of contextual representations across neighboring windows. Lastly, we find that cross-window regularization that aligns window-specific *CLS* tokens contributes notably to model performance. This result indicates that the model benefits from coherence of representations in *CLS* tokens that are averaged across windows to implement classification. Overall, we find that the BolT model including all of its design elements yields the highest performance among all variants.

Compared to the non-windowed variant with *CLS*, the ablated variant with windowing and local *CLS* tokens yields a notable performance improvement on the AAL atlas, albeit it shows relatively stable performance on the Schaefer atlas (Table 3.1). Since the measured fMRI data and modeling procedures were identical for the two atlases, this difference is best attributed to an interaction between window size and the temporal characteristics of ROI responses. In theory, a larger W can be suited to analyze relatively slow varying responses, and a smaller W can be suited to analyze relatively fast varying responses. A power spectral density analysis shows that ROI responses based on the AAL atlas that groups voxels based on anatomical proximity carry higher energy at high temporal frequencies, whereas responses based on the Schaefer atlas that groups voxels based on functional similarity carry higher energy at low temporal frequencies (not reported). In turn, we observe that the performance benefits from split time windows are maximized at $W=20$ for the AAL atlas, and at $W=200$ for the Schaefer atlas (not reported). Note, however, that BolT does not only use basic windowing

as in the ablated variant, but it also leverages additional window-related design elements including cross-attention between base and fringe tokens, token fusion and cross-window regularization. These elements promote information exchange across separate time windows, increasing the effective temporal receptive field for each window. As such, a small W can serve to sensitively analyze a broad range of responses. Table 3.2 lists performance metrics for BolT under varying window sizes. We find that optimal or near-optimal performance is attained at a compact window size of $W=20$ commonly for both atlases. This result suggests that window-related design elements in BolT collectively introduce a degree of reliability against varying temporal frequency characteristics of BOLD responses.

Next, we examined the efficacy of the explanatory technique that computes an importance weight for each BOLD token (see Fig. 3.3). This importance weight is supposed to reflect the degree of discriminative information captured by the token for the respective detection task. We reasoned that if the explanatory technique computes reasonable weights, significant detection should be possible based on a subset of highly important tokens. To test this prediction, BOLD tokens in the time series were ordered according to their importance weights. ROI definitions based on the Schaefer atlas were used for this analysis, since they yielded better performance in BolT. Logistic-regression models were then built for the same detection task given as input a subset of five consecutive tokens [153, 154]. Table 3.3 lists detection performance based on the most important subset of tokens against that based on a subset of randomly selected tokens. While random tokens perform near chance level, important tokens achieve substantially higher performance. We also reasoned that detection performance should scale with the overall importance of the selected token subset. To examine this issue, separate logistic-regression models were built while the overall importance of the token subset was systematically reduced (Fig. 3.4). Detection performance elevates near-monotonically with increasing levels of token importance. Taken together, these results indicate that the importance weights returned by the explanatory technique closely reflect the contribution of individual tokens to model decisions.

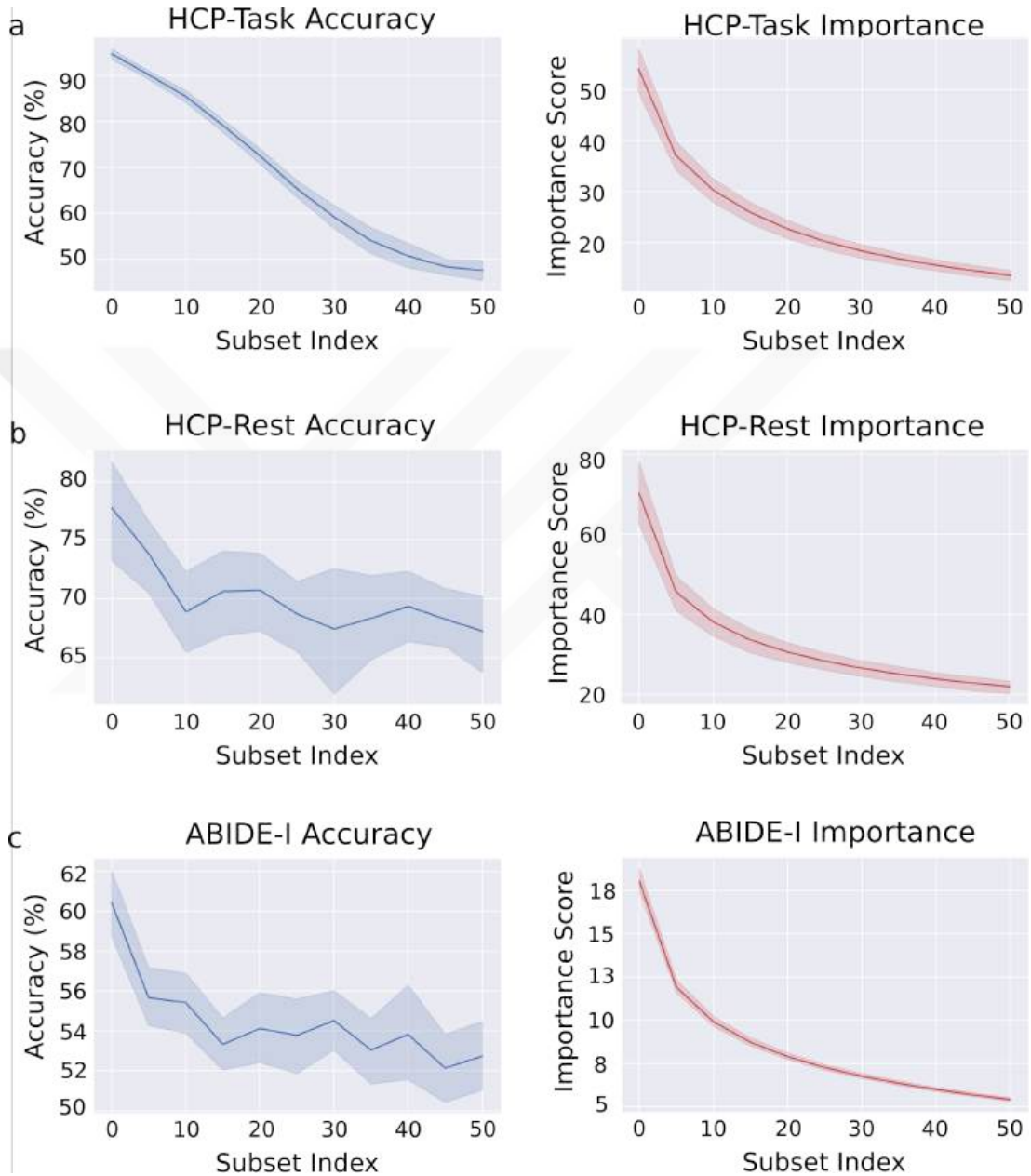


Figure 3.4: Accuracy of logistic-regression models receiving as input a subset of five important tokens, while the total importance of the subset is systematically varied. Tokens were ordered according to their importance weights, different subsets of five consecutive tokens in the ordered list were selected. Results are shown for varying subset index for (a) HCP-Task, (b) HCP-Rest, (c) ABIDE-I. Subset index refers to the offset within the ordered list for the selected subset. Importance score was taken as the average importance weight of tokens in a given subset normalized by the minimum importance weight within the time series. As such, total importance of the subset and detection accuracy show a general decrease with increasing subset index.

Table 3.4: Performance of competing methods and BolT using Schaefer atlas for gender detection on the HCP-Rest dataset, task detection on the HCP-Task dataset and disease detection on the ABIDE-I dataset. Metrics are reported as mean $\pm$ std across test folds. Boldface indicates the top-performing model in terms of each metric in individual classification tasks.

	HCP-Rest (Schaefer)				HCP-Task (Schaefer)				ABIDE-I (Schaefer)			
	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)
SVM	75.92 ± 4.78	65.13 ± 7.41	78.63 ± 6.27	85.34 ± 4.59	95.57 ± 0.87	95.58 ± 0.88	95.61 ± 0.86	99.78 ± 0.07	65.72 ± 4.11	53.97 ± 7.33	66.26 ± 6.68	71.33 ± 4.62
BrainNetCNN	84.16 ± 3.73	82.97 ± 6.63	82.39 ± 3.04	90.95 ± 3.85	96.42 ± 0.82	96.42 ± 0.87	96.52 ± 0.82	99.81 ± 0.06	66.78 ± 3.90	63.82 ± 8.73	65.50 ± 4.63	73.15 ± 5.62
BrainGNN	79.14 ± 3.90	78.96 ± 5.43	76.89 ± 6.19	86.31 ± 2.51	93.10 ± 1.40	93.11 ± 1.43	93.80 ± 1.07	99.67 ± 0.12	59.87 ± 5.80	47.86 ± 12.21	57.79 ± 6.50	63.86 ± 6.63
STAGIN	82.51 ± 3.69	84.16 ± 5.61	79.20 ± 4.71	87.63 ± 3.53	99.27 ± 0.43	99.26 ± 0.43	99.27 ± 0.42	99.98 ± 0.01	61.70 ± 3.56	41.69 ± 7.37	63.63 ± 7.03	64.71 ± 6.02
LSTM	81.59 ± 4.03	82.16 ± 3.00	78.99 ± 6.37	90.46 ± 2.37	98.35 ± 0.65	98.34 ± 0.65	98.41 ± 0.58	99.94 ± 0.04	64.55 ± 5.41	57.14 ± 16.20	62.85 ± 6.98	68.88 ± 5.87
CNN-LSTM	80.77 ± 3.83	79.57 ± 9.34	80.38 ± 9.27	88.47 ± 3.39	99.04 ± 0.59	99.04 ± 0.59	99.07 0.56	99.96 ± 0.05	65.49 ± 5.70	57.31 ± 9.40	64.79 ± 7.57	71.40 ± 5.41
GC-LSTM	83.99 ± 4.85	76.90 ± 16.09	87.64 ± 5.14	93.85 ± 1.35	98.16 ± 0.59	98.16 ± 0.59	98.20 ± 0.59	99.96 ± 0.02	62.77 ± 5.46	60.24 ± 19.56	62.11 ± 10.96	68.28 ± 4.84
SwinT	79.41 ± 2.49	77.35 ± 4.54	77.67 ± 3.63	87.49 ± 1.71	99.54 ± 0.27	99.54 ± 0.27	99.55 ± 0.27	99.99 ± 0.00	68.56 ± 4.74	60.75 ± 7.60	68.27 ± 6.56	74.13 ± 4.15
Longformer	83.24 ± 3.34	79.15 ± 8.49	84.03 ± 5.45	92.41 ± 2.81	99.46 ± 0.39	99.46 ± 0.38	99.47 ± 0.38	99.99 ± 0.00	67.85 ± 4.58	57.15 ± 15.79	70.09 ± 7.13	74.87 ± 3.99
BaND	83.61 ± 4.03	84.54 ± 6.41	80.87 ± 5.42	92.62 ± 2.20	99.24 ± 0.46	99.24 ± 0.46	99.26 ± 0.43	99.98 ± 0.01	65.48 ± 3.04	58.07 ± 7.93	64.47 ± 4.76	72.10 ± 4.38
TFF	87.36 ± 3.68	84.17 ± 4.21	87.89 ± 5.09	94.79 ± 2.39	99.08 ± 0.25	99.08 ± 0.25	99.11 ± 0.24	99.98 ± 0.02	66.73 ± 5.33	42.71 ± 16.49	78.93 ± 10.19	75.44 ± 3.94
IFT-Net	82.97 ± 2.84	79.76 ± 10.28	83.72 ± 7.27	93.12 ± 2.05	97.58 ± 3.62	97.58 ± 3.62	98.08 ± 2.59	99.91 ± 0.16	61.88 ± 5.52	51.59 ± 23.88	63.82 ± 9.42	68.50 ± 6.44
HATNet	85.72 ± 2.64	84.38 ± 5.89	84.57 ± 3.68	93.66 ± 2.26	99.37 ± 0.33	99.36 ± 0.33	99.38 ± 0.32	99.99 ± 0.00	64.66 ± 4.53	57.67 ± 6.88	62.99 ± 5.76	68.92 ± 4.95
BolT	91.85 ± 3.05	90.58 ± 4.97	91.51 ± 3.07	97.35 ± 1.06	99.66 ± 0.35	99.66 ± 0.35	99.67 ± 0.34	99.99 ± 0.00	71.28 ± 4.62	64.85 ± 7.94	71.32 ± 7.35	77.56 ± 3.44

Table 3.5: Performance of competing methods and BolT using AAL atlas for gender detection on the HCP-Rest dataset, task detection on the HCP-Task dataset and disease detection on the ABIDE-I dataset. Metrics are reported as mean $\pm$ std across test folds. Boldface indicates the top-performing model in terms of each metric in individual classification tasks.

	HCP-Rest (AAL)				HCP-Task (AAL)				ABIDE-I (AAL)			
	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)	Acc.(%)	Rec.(%)	Prec.(%)	AUC (%)	Acc.(%)	Rec.(%)	Prec.(%)	AUC(%)
SVM	71.71 ± 3.76	65.40 ± 11.86	70.55 ± 3.28	78.30 ± 4.09	88.11 ± 1.46	88.14 ± 1.46	88.25 ± 1.38	98.54 ± 0.31	65.60 ± 3.14	55.00 ± 7.40	64.10 ± 5.77	71.16 ± 4.18
BrainNetCNN	71.16 ± 4.91	67.40 ± 9.21	69.20 ± 6.47	78.99 ± 5.12	90.56 ± 1.06	90.58 ± 1.07	90.73 ± 1.04	99.04 ± 0.27	64.01 ± 6.04	60.24 ± 9.33	61.86 ± 6.75	69.79 ± 6.16
BrainGNN	69.79 ± 4.59	72.40 ± 9.70	65.25 ± 4.36	77.85 ± 4.59	79.73 ± 5.92	79.85 ± 5.88	81.72 ± 4.77	97.05 ± 1.29	61.40 ± 4.66	58.30 ± 10.86	58.40 ± 6.74	65.68 ± 5.61
STAGIN	76.18 ± 3.10	72.20 ± 9.22	75.00 ± 3.75	83.07 ± 3.15	98.87 ± 0.60	98.86 ± 0.61	98.91 ± 0.57	99.95 ± 0.03	61.52 ± 3.49	52.69 ± 7.38	60.12 ± 4.94	66.68 ± 4.36
LSTM	73.25 ± 4.48	67.19 ± 8.44	73.10 ± 7.59	81.96 ± 2.90	96.96 ± 0.69	96.97 ± 0.69	97.06 ± 0.64	99.88 ± 0.06	63.06 ± 3.96	45.81 ± 22.76	64.05 ± 13.93	70.25 ± 3.85
CNN-LSTM	74.81 ± 3.15	66.40 ± 9.20	76.25 ± 5.28	82.88 ± 3.30	97.77 ± 0.60	97.76 ± 0.61	97.83 ± 0.57	99.91 ± 0.05	63.65 ± 5.42	49.47 ± 17.04	66.49 ± 11.86	68.78 ± 4.13
GC-LSTM	77.27 ± 5.29	77.79 ± 14.57	77.61 ± 11.50	89.59 ± 1.93	93.14 ± 1.39	93.14 ± 1.39	93.74 ± 1.09	99.57 ± 0.14	60.05 ± 5.36	48.27 ± 19.80	61.04 ± 9.04	64.67 ± 6.90
SwinT	78.37 ± 4.00	76.60 ± 6.69	76.14 ± 4.07	84.84 ± 3.58	99.23 ± 0.38	99.23 ± 0.38	99.25 ± 0.36	99.99 ± 0.01	66.78 ± 4.25	60.48 ± 6.38	65.56 ± 5.57	72.92 ± 4.30
Longformer	76.28 ± 4.30	63.00 ± 13.94	83.53 ± 9.68	87.80 ± 3.10	99.11 ± 0.42	99.11 ± 0.43	99.13 ± 0.41	99.99 ± 0.00	64.77 ± 4.71	58.69 ± 18.18	64.99 ± 9.42	71.42 ± 4.70
BaND	78.55 ± 4.12	70.40 ± 8.38	80.46 ± 5.18	87.39 ± 2.84	98.16 ± 0.45	98.16 ± 0.45	98.20 ± 0.41	99.93 ± 0.03	63.12 ± 3.60	45.18 ± 12.53	65.25 ± 6.33	68.65 ± 3.61
TFF	82.57 ± 4.05	81.80 ± 8.59	81.34 ± 7.41	91.14 ± 2.98	97.43 ± 1.00	97.43 ± 1.00	97.55 ± 0.90	99.90 ± 0.07	65.84 ± 3.87	46.30 ± 15.83	73.30 ± 8.86	74.14 ± 4.16
IFT-Net	77.72 ± 4.85	76.00 ± 11.76	76.57 ± 7.28	86.69 ± 2.62	95.22 ± 5.58	95.17 ± 5.67	96.19 ± 3.80	99.87 ± 0.19	58.26 ± 4.72	32.11 ± 24.85	54.23 ± 25.90	64.99 ± 4.64
HATNet	78.37 ± 2.15	74.60 ± 5.58	77.23 ± 1.54	87.43 ± 2.96	97.97 ± 0.35	97.96 ± 0.35	98.02 ± 0.32	99.96 ± 0.02	61.16 ± 5.64	54.59 ± 7.15	58.79 ± 6.55	66.52 ± 5.10
BolT	87.31 ± 2.69	86.99 ± 4.49	85.65 ± 4.01	94.29 ± 2.05	99.52 ± 0.39	99.52 ± 0.40	99.54 ± 0.38	99.99 ± 0.00	68.14 ± 2.81	60.49 ± 4.22	67.52 ± 4.07	74.30 ± 3.69

3.5.2 Comparative demonstration of BolT

We demonstrated BolT for three main tasks in fMRI analysis: gender detection on HCP-Rest, cognitive task detection on HCP-Task, and disease detection on ABIDE-I datasets. BolT was demonstrated against state-of-the-art traditional (SVM), CNN (BrainNetCNN), GNN (BrainGNN, STAGIN), RNN (LSTM, CNN-LSTM, GC-LSTM), and transformer (SwinT, Longformer, BaND, TFF, IFT-Net, HATNet) baselines. Demonstrations were performed using ROI definitions extracted via two different brain atlases. Performance metrics for competing methods for Schaefer atlas are listed in Table 3.4, and those for AAL atlas are listed in Table 3.5. For each detection task and based on each atlas, BolT outperforms all competing methods in each metric ($p < 0.05$, Wilcoxon signed-rank test), except for TFF that offers higher precision on ABIDE-I, SwinT that offers similar recall on ABIDE-I (AAL atlas), and SwinT, Longformer, and HATNet that offer similar AUC on HCP-Task. On average across atlases in gender detection, BolT improves (accuracy, recall, precision, AUC) by (8.39, 11.13, 7.41, 5.87)% over transformer baselines, (10.96, 13.78, 9.58, 7.95)% over RNN baselines, (12.67, 11.85, 14.49, 12.10)% over GNN baselines, (11.92, 13.59, 12.78, 10.84)% over the CNN baseline, and (15.76, 23.51, 13.99, 14.00)% over the traditional baseline. In cognitive task detection, BolT achieves improvements of (1.14, 1.14, 1.00, 0.03)% over transformer baselines, (2.35, 2.35, 2.21, 0.12)% over RNN baselines, (6.84, 6.82, 6.18, 0.82)% over GNN baselines, (6.09, 6.09, 5.98, 0.56)% over the CNN baseline, and (7.75, 7.73, 7.67, 0.82)% over the traditional baseline. Finally, in disease detection, BolT achieves improvements of (5.11, 10.56, -, 4.88)% over transformer baselines, (6.44, 9.63, 5.86, 7.22)% over RNN baselines, (8.58, 12.53, 9.43, 10.69)% over GNN baselines, (4.31, 0.64, 5.73, 4.46)% over the CNN baseline, and (4.05, 8.15, 4.23, 4.68)% over the traditional baseline. Taken together, these results indicate that BolT enables significant performance benefits in detection tasks over prior traditional and DL methods.

In general, we observe that DL models yield superior performance to the traditional SVM baseline on HCP-Rest and HCP-Task, whereas SVM outperforms CNN, GNN, RNN, and a subset of transformer baselines in disease detection

on ABIDE-I. Note that HCP-Rest and HCP-Task were acquired using relatively standardized protocols and scanner hardware in a compact set of imaging sites. In contrast, ABIDE-I was curated by aggregating data from a larger number of sites with more substantial variations in imaging protocols and hardware. In turn, the resultant data heterogeneity can limit generalization performance for DL methods with relatively high complexity, while the simpler SVM method starts performing competitively. That said, we observe that windowed transformer models including BolT still outperform SVM in this case, implying a degree of reliability against data heterogeneity due to the generalization capabilities of self-attention operators combined with local sensitivity from split time windows. We also observe that all competing methods yield notably higher performance on HCP-Task, compared to HCP-Rest and ABIDE-I. This is expected as detecting divergent cognitive tasks from task-based fMRI scans that elicit responses in largely non-overlapping brain networks is relatively easier compared to detection tasks on resting-state fMRI scans. Here, we preferred to report HCP-Task since it is a highly relevant, benchmark dataset that is frequently reported in methodology studies on task-based fMRI analysis. Yet, future studies are warranted to examine the utility of BolT in detecting BOLD-response differences among more similar cognitive tasks driving partly overlapping brain networks.

3.5.3 Explainability of BolT

To interpret the spatio-temporal patterns of brain activation that contribute to BolT’s decisions, we employed the explanatory technique to calculate token importance weights. Importance weights for each cognitive task in HCP-Task are shown in Fig. 3.3. Landmark time points of high importance closely align with transitions in the temporal structure of task variables following an offset due to hemodynamic delay. For instance, periods of target maintenance following target appearance are attributed high importance in the working memory task, corresponding to abrupt changes in activation [154].

We then leveraged the landmark time points to identify brain regions critical

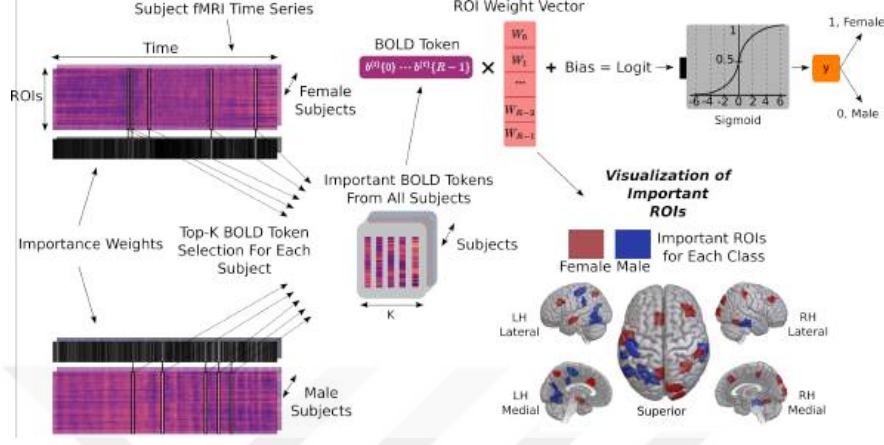


Figure 3.5: Landmark time points (i.e. BOLD tokens) selected by BolT were used to identify brain regions critical for gender detection in HCP-Rest. A collection of $K=5$ tokens were retrieved from each subject, characterizing responses across R ROIs. Next, a logistic regression model was trained to map the tokens in landmark time points onto the associated output class. Model weights reflect each ROI’s contribution to the classification decision. For each class, the top 2 percent of most influential ROIs were visualized (i.e. female in red color, male in blue color).

for the detection tasks. To do this, a logistic-regression model was trained on top-five most important BOLD tokens, and model weights were taken to reflect the importance of individual ROIs for task performance [155]. As shown in Fig. 3.5 for gender detection, we find important ROIs across the attention and somatosensory networks in male subjects, and ROIs in prefrontal/frontal cortices and default mode network (DMN) in female subjects. This is consistent with previous reports on stronger FC features across sensorimotor cortices in males and across DMN in females [156, 157, 158]. We further find important ROIs in visual networks for both genders. This result is aligned with a recent report suggesting that responses in visual regions might implicitly represent gender-discriminating information [128]. As shown in Fig. 3.6 for task detection, brain regions implicated with the target task are attributed high importance (e.g., sensorimotor regions in the Motor task, temporal regions in the Language task). As shown in Fig. 3.7 for ASD detection, we find important ROIs in healthy controls across the frontal-parietal network (FPN), thought to mediate goal-oriented, cognitively demanding behavior [159]. In contrast, ASD patients manifest important ROIs across DMN, with commonly reported over-activation in ASD [160, 137, 161, 21].

Taken together, these results indicate that BolT effectively captures task-relevant patterns of brain activation in both normal and disease states.

3.6 Discussion

Here, we introduced a transformer architecture that efficiently captures local-to-global representations of time series to perform detection tasks based on fMRI scans. The proposed architecture learns latent representations of fMRI data via a novel fused window attention mechanism that incorporates long-range context with linear complexity in terms of scan length. Detection is then performed based on learned high-level classification tokens regularized across time windows. Demonstrations were performed on resting-state and task-based fMRI data with superior performance against state-of-the-art baselines including convolutional, graph and transformer models.

In this study, we primarily built classification models with categorical output variables for gender, cognitive task, and disease. To improve classification performance, learnable *CLS* tokens were included that provide a condensed high-level representation of corresponding time windows. Note that the human brain does not only represent categorical variables, but it is also assumed to carry information regarding continuous stimulus or task features [45]. To analyze cortical representations of such continuous features, BolT can be adapted to instead build regression models [65, 162, 163]. To do this, latent representations of BOLD tokens in downstream layers of BolT can be coupled with linear or ReLU activation functions. This can enable BolT to decode continuous stimulus or task variables from fMRI scans.

A common approach to interrogate brain function in computational neuroimaging studies is to build decoding models that predict stimulus or task features given as input measured brain activations [164, 165]. Following this framework, here we used BolT to build decoding models that detect external features based on measured BOLD responses. An alternative framework rests on analysis of brain

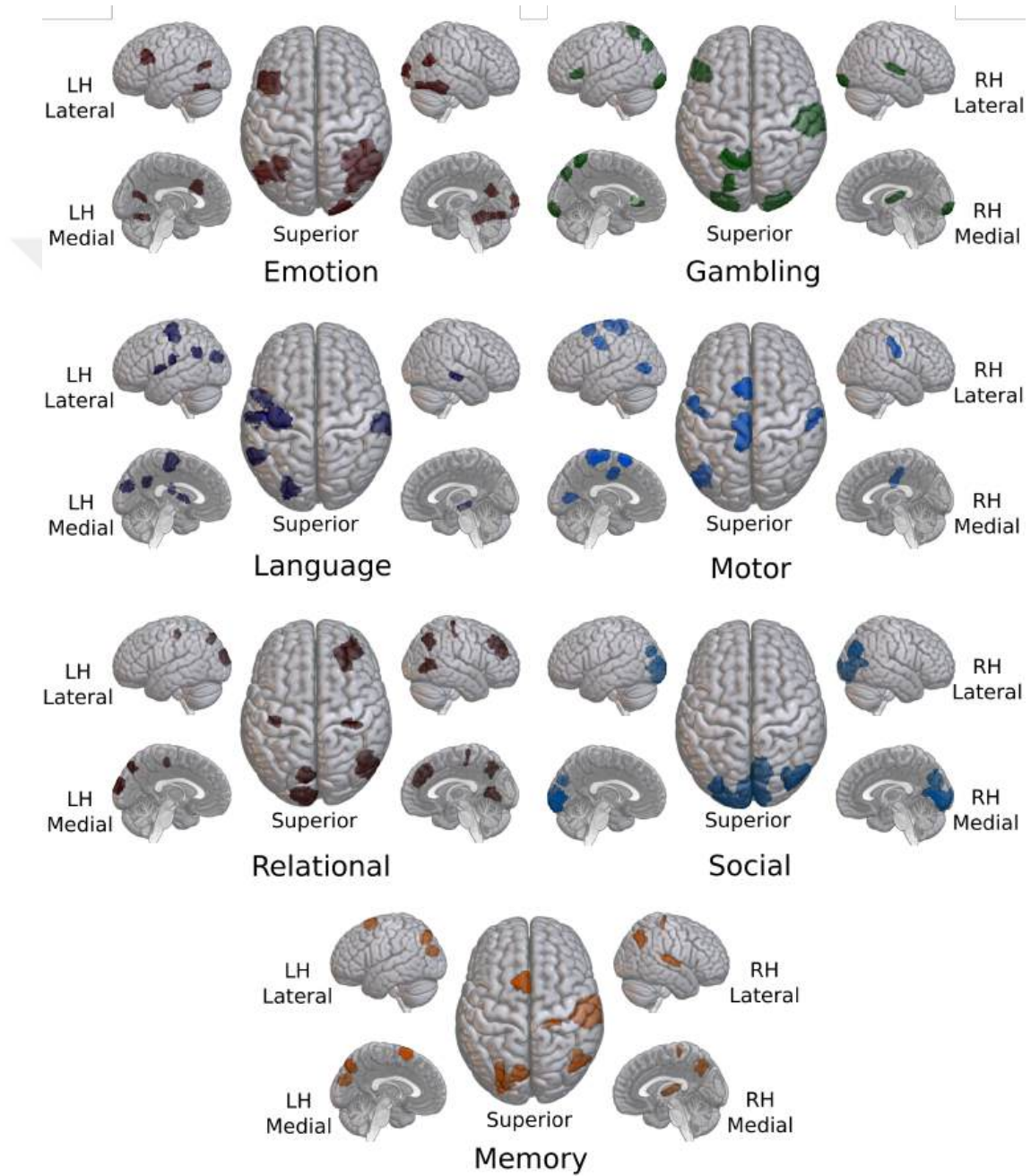


Figure 3.6: Critical ROIs for cognitive task detection in HCP-Task. For each task, the top 2 percent of most influential ROIs were visualized. Elevated BOLD responses in highlighted ROIs imply the presence of the associated task.

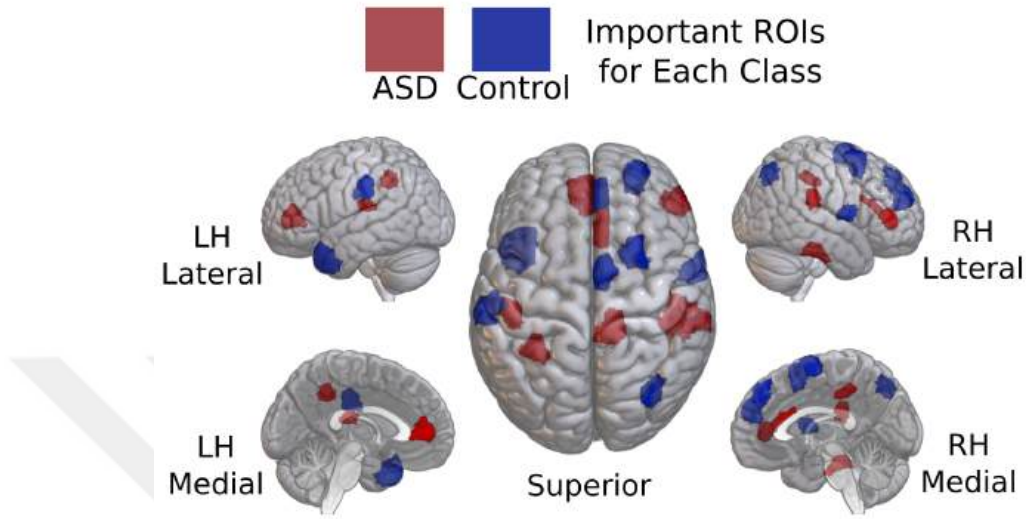


Figure 3.7: Critical ROIs for ASD detection in ABIDE-I. The top 2 percent of most influential ROIs were visualized for ASD patients and healthy controls. Elevated BOLD responses in red ROIs imply the presence of the ASD condition, whereas elevated responses in blue ROIs imply a healthy condition.

function by building encoding models that predict brain activations given as input stimulus/task features [166, 50, 48, 167, 168]. In cognitive neuroimaging studies, the experimental time course for the stimulus and/or cognitive task can be taken as input to BolT, and voxel-wise regression models can be built to estimate measured BOLD responses. It remains important future work to assess the efficacy of BolT in training encoding models.

Literature suggests that resting-state fMRI scans carry idiosyncratic information regarding disease progression in neurodevelopmental disorders [169, 170]. Based on this literature, we considered ASD detection using solely information from resting-state fMRI scans. Recent studies suggest that auxiliary information on patient demographics or scan protocols might help facilitate disease detection [171]. Moreover, some neurological diseases such as Alzheimer’s or dementia have complementary imaging signatures in other modalities such as structural or diffusion-weighted MRI [172]. Thus, it is reasonable to expect that disease detection performance with BolT can be further improved by incorporating auxiliary information as well as additional imaging modalities. Auxiliary information can

be integrated via bypass channels near the output layers of BolT, whereas additional imaging modalities can be incorporated as added input channels alongside fMRI data.

As commonly practiced in many fMRI studies, here we first normalized each subject’s brain volume onto an anatomical template, and then used an anatomical atlas to define brain ROIs. Average BOLD responses in individual ROIs were then provided as input to BolT. Note that this approach ensures relatively consistent and comprehensive ROI definitions across subjects, permitting analyses in brain regions that do not have well established functional-localization procedures [173]. Yet, spatial registration to a common template involves a potentially lossy transformation of fMRI data. Such losses can be mitigated by defining ROIs in the brain spaces of individual subjects as opposed to a template. To do this, the registration transform between the subject and template brain spaces can be estimated. ROI boundaries in the template brain space can then be backprojected onto the individual subject brain space by inverting the estimated transformation [43, 48, 174]. Alternatively, a CNN model can also be incorporated in BolT to perform spatial encoding of volumetric MRI data prior to processing with the transformer blocks [98, 99].

Here we trained all competing models from scratch on fMRI data from several hundred subjects for each detection task. Given their relatively higher complexity against convolutional models, transformers are generally considered to require substantial datasets for successful learning [145, 18]. In applications where only compact datasets are available, pre-training and transfer learning procedures can be adopted to initialize the network weights in transformer architectures [175, 176]. Reliable augmentation via image synthesis based on advanced procedures such as diffusion modeling can also help alleviate data scarcity [177, ?, 178]. Alternatively, complexity of self-attention modules can be mitigated by replacing regular dot-product attention operators with efficient kernelized operators [136]. Federated learning across multiple institutions might facilitate learning from large, diverse datasets without introducing privacy risks [27, ?]. Lastly, unsupervised learning strategies can also be adopted to permit training on partially labeled fMRI datasets from a larger subject cohort [99, 29, 179]. A systematic

exploration of the data efficiency of BolT against competing models remains an important topic for future research.



Chapter 4

A plug-in graph neural network to boost temporal sensitivity in fMRI analysis

The content of this chapter reflects the study described in the following publications:

- I. Sivgin, **H. A. Bedel**, S. Ozturk, and T. Çukur, "A plug-in graph neural network to boost temporal sensitivity in fMRI analysis," IEEE J. Biomed. Health Inform., 2024. Conference)

4.1 Introduction

The human brain comprises functional networks composed of multiple brain regions that interactively process information to mediate cognitive processes [45].

In turn, correlated activity among brain regions within individual functional networks has been associated with unique mental states [180, 181, 182, 183]. Functional MRI (fMRI) is a powerful modality to examine networks as it can non-invasively measure whole-brain blood-oxygen-level-dependent (BOLD) signals consequent to neural activity at high spatio-temporal resolution [184, 182, 63]. In fMRI studies, networks are commonly assessed via functional connectivity (FC) measures that reflect similarity of BOLD signals among brain regions [139, 185, 63]. The traditional approach to map FC measures onto mental states is then based on conventional methods such as logistic regression and support vector machines (SVM) [186, 187, 188, 189]. Unfortunately, conventional methods often yield poor capture of intricate information patterns in whole-brain fMRI time series [?].

In recent years, the success of deep learning (DL) models at exploring features in high-dimensional datasets has motivated their adoption for fMRI analysis as an alternative to conventional methods [190, 91, 126, 86, 191]. Earlier attempts in this domain have proposed shallow multi-layer perceptron (MLP) [116, 192] and Boltzmann machine (BM) models [193, 190, 194]. Later studies have adopted deeper architectures based on convolutional neural network (CNN) [123, 91, 124], graph neural network (GNN) [195, 126, 196, 139, 197, 198, 199, 127], and transformer [200, 99, 98, 201] models for improved performance. These models construct a set of nodes corresponding to brain regions defined based on an atlas [151, 202, 203, 199], and receive input features at these nodes based on the FC strength among brain regions [204, 126, 205]. A pervasive approach has been to employ static FC features derived from summary correlation measures across the entire fMRI scan [206, 126]. Yet, this approach is often insufficiently sensitive to the dynamic inter-regional interactions during resting-state or cognitive tasks [107]. While alternative strategies have recently been proposed to capture the temporal variability in FC features, these methods commonly consider instantaneous correlations¹ across local time windows within the time series [128, 207, 208]. As such, they lack explicit mechanisms to capture delayed

¹**Instantaneous correlation:** Cross-correlation between BOLD signals of a pair of brain regions under zero time delay.

correlations² eminent in fMRI data due to hierarchical cognitive processing or hemodynamic lags in BOLD signals [50].

Here we introduce a novel plug-in graphical neural network, GraphCorr, that provides enhanced input features to baseline classification models so as to boost their sensitivity to dynamic inter-regional interactions in fMRI data. To capture instantaneous interactions, GraphCorr leverages a novel node embedder module based on a transformer encoder that computes hierarchical embeddings of windowed BOLD signals across the time series. To capture lagged interactions, GraphCorr employs a novel lag filter module that computes nonlinear features of cross-correlation between pairs of nodes across a range of time delays. The graph is initialized with node features taken as hierarchical embeddings, and edge weights taken as cross-correlation features. Afterwards, a message passing algorithm is used to compute enhanced node features that account for dynamic, lagged inter-regional interactions, while maintaining comparable feature dimensionality to static FC features.

To demonstrate GraphCorr, we conduct comprehensive experiments on two benchmark tasks and three public datasets frequently reported in the literature. Sex is assumed to be an important biological variable driving the functional organization of the brain in normal and disease states, and recent studies report gender-specific differences in functional connectivity among brain regions [156, 209, 204]. Thus, we first examine gender detection on resting-state scans in the HCP-Rest dataset from Human Connectome Project [103] and on movie-watching fMRI scans in the ID1000 dataset from Amsterdam Open MRI Collection [210]. Functional connectivity differences are not exclusive to resting state, but they have also been reported during intentful execution of various cognitive tasks [167, 127, 128]. Thus, we also examine cognitive-task detection on task-based fMRI scans in the HCP-Task dataset [103].

GraphCorr is used as a plug-in to augment state-of-the-art learning-based

²**Delayed correlation:** Cross-correlation between BOLD signals of a pair of brain regions under non-zero time delay.

fMRI classifiers including graphical (SAGE [211], GCN [197]), and convolutional (BrainNetCNN [91]) baselines. We find that, for each baseline model, the GraphCorr-augmented variant significantly outperforms the vanilla variant as well as variants augmented with dynamic FC features or with gated recurrent units (GRU). We further devise an explanatory approach for GraphCorr to interpret the brain regions that most significantly contribute to classification decisions. We show that GraphCorr offers interpretations that are closely aligned with prominent neuroscientific findings from the literature.

4.2 Related Work

Cognitive processes in the human brain elicit broadly distributed response patterns spanning across multiple brain regions [106], which can be analyzed to classify stimulus or task variables [212, 43, 213]. Many analysis methods for fMRI rely on feature selection procedures to cope with the intrinsically high dimensionality of fMRI data [107, 82, 113]. Arguably, FC measures among brain regions have been the most commonly used feature set [116, 91, 204, 126, 214]. Previous studies have reported that external variables or disease states can be detected given FC features of individual subjects under resting state [215, 216, 217], cognitive tasks [186, 126], or both [189]. Given their success in fMRI analysis, FC features have also been adopted in recent DL methods [218, 219, 220]. Most commonly, static FC features have been employed that take FC between a pair of regions as the aggregate correlation of their BOLD signals across the entire scan. To extract latent representations of FC features, initial studies have proposed either shallow fully-connected architectures [193, 190, 116, 192], or deep convolutional architectures [124, 91]. Later studies have considered GNN models given their natural fit to analyzing fMRI data that follows an intrinsic graph-like connectivity structure [221, 222, 223, 224, 126, 127, 225]. While these DL methods have enabled substantial improvements over traditional methods, analyses based on static FC features can yield suboptimal sensitivity to fine-grained temporal information [187, 124].

Several groups of strategies have been proposed to incorporate time-varying features into DL-based fMRI analysis. A first group pre-computes FC features over moving windows across the time series based on standard correlation measures, and concatenate them across windows [226, 102, 139, 207, 208, 227]. While these dynamic FC features carry enhanced temporal information, their dimensionality grows with the number of time windows and can undesirably increase complexity in baseline classification models. A second group instead uses voxel-level BOLD signals spatially encoded via a CNN module as model input, and employ RNN or transformer models to extract time-varying information [99, 128]. Yet, CNN modules based on voxel-level inputs can be difficult to train from scratch under limited data regimes. A third group retains static FC features as their input, albeit augments them with dynamic features that RNN modules capture from BOLD signals [128]. Besides elevated model complexity, these methods can suffer from limitations of RNNs in terms of vanishing/exploding gradients [228, 100, 229]. Importantly, these previous methods primarily focus on temporal variations in instantaneous correlations, while neglecting delayed inter-regional correlations [50].

Here, we propose to improve the classification performance of baseline models via a novel plug-in, GraphCorr, that extracts enhanced features with fine-grained temporal information from fMRI data. The proposed plug-in embodies several unique technical attributes of potential value for fMRI analysis. Unlike methods based on static FC features [91, 126, 230], GraphCorr leverages dynamic FC features to capture the temporal variability in connectivity among brain regions. Unlike methods that receive multiple sets of dynamic FC features across separate time windows [75, 128, 139, 231], GraphCorr leverages a message passing algorithm formulated on a graph to fuse the dynamic FC features it captures across windows to reduce feature dimensionality while maintaining fine-grained temporal information, thereby it avoids undesirable increases in complexity of baseline classification models. Unlike methods that employ recurrent architectures [128], it leverages a transformer encoder on dynamic FC features that enables efficient parallel processing. Unlike methods that solely focus on instantaneous correlations [91], it adopts an explicit lag filter mechanism to learn delayed cross-correlations.

These advances enable GraphCorr to enhance the level of temporal information available in the features input to baseline classification models.

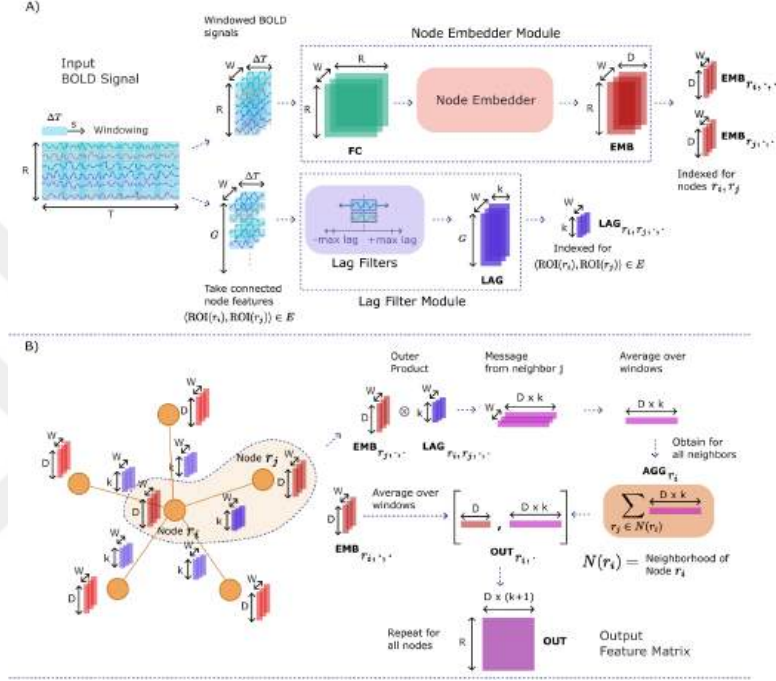


Figure 4.1: Overview of GraphCorr. A) GraphCorr utilizes two modules in parallel to extract dynamic, lagged features of inter-regional correlations across the brain. The node embedder module receives as input time-windowed BOLD signals, and uses a transformer encoder to compute node embeddings of dynamic FC features $\mathbf{EMB} \in \mathbb{R}^{R \times D \times W}$ (R : number of ROIs, D : embedding dimensionality, W : number of windows). The lag filter module also receives as input time-windowed BOLD signals, albeit it computes lag activations due to cross-correlation between node pairs across a range of time delays $\mathbf{LAG} \in \mathbb{R}^{R \times R \times W \times k}$ (k : number of lag filters). B) To consolidate the extracted feature sets on a graph, node embeddings are taken as node features and lag activations are taken as edge weights. A message passing algorithm is then run on the graph to produce enhanced FC features in an output feature matrix, $\mathbf{OUT} \in \mathbb{R}^{R \times (Dk+D)}$.

4.3 Methods

Analysis procedures for fMRI time series typically start by defining a collection of R regions-of-interest (ROI) across the brain based on an atlas [126, 128]. Voxel-level BOLD signals within each ROI are then averaged to derive ROI-level signals,

resulting in $\mathbf{B} \in \mathbb{R}^{R \times T}$ as the matrix of BOLD signals where T denotes the number of time frames, and $\mathbb{R}$ denotes the real set. Static FC features are conventionally computed based on Pearson’s correlation coefficient between the rows of this matrix: $\mathbf{sFC}_{r_i, r_j} = \text{Corr}(\mathbf{B}_{r_i, \cdot}, \mathbf{B}_{r_j, \cdot})$, where $\mathbf{sFC} \in \mathbb{R}^{R \times R}$ and $r_i, r_j \in 1, 2, \dots, R$ are ROI indices. While previous traditional and learning-based methods commonly operate on static FC features, here we propose to extract latent FC features with enhanced temporal information based on a novel GNN plug-in, and to use these enhanced features to improve the performance of baseline models³. Formulated on a graph, GraphCorr leverages node embedder and lag filter modules to capture dynamic, lagged correlations in BOLD signals, and performs message passing on the graph to learn enhanced features (Fig. 4.1). The methodological components and procedures are described below. Code for GraphCorr will be shared at <https://github.com/icon-lab/GraphCorr> upon publication.

4.3.1 Graph formation

As its learning substrate, GraphCorr first forms a binary graph $G(N, E)$ with a set of nodes N and a corresponding set of edges E that reflect connections between pairs of nodes. The node set $N = \{\text{ROI}(r_i) | r_i = 1, \dots, R\}$ includes a total of R ROIs defined according to the atlas. Meanwhile, the binary edge set is taken as $E = \{\langle \text{ROI}(r_i), \text{ROI}(r_j) \rangle | r_i = 1, \dots, R; r_j \in \mathcal{N}(r_i)\}$ where $\langle \cdot, \cdot \rangle$ denotes a connection between two nodes, and $\mathcal{N}(r_i)$ denotes the neighborhood of nodes connected to the r_i -th node. Binary edges are derived by thresholding the elements in $\mathbf{sFC}$ to retain connections for the top $z\%$ percentile of correlation coefficients (excluding self-connections), resulting in a total of G edges [126]. An initial tensor of node features $\mathbf{F} = \{f_{r_i} | r_i = 1, \dots, R\}$ are defined based on time-windowed BOLD signals to capture local dynamics in the fMRI time series. For this purpose, the time series for each ROI containing T time frames is split into W windows of size ΔT and stride value s :

$$W = \lfloor \frac{T - \Delta T}{s} \rfloor, \quad (4.1)$$

³see [232] for a preliminary version of this work presented at SIU 2022

For a given node $\text{ROI}(r_i)$, this split time series is reformatted into a matrix of windowed BOLD signals, $f_{r_i} \in \mathbb{R}^{\Delta T \times W}$. The matrices for individual nodes are concatenated along the ROI dimension to yield the feature tensor $\mathbf{F} \in \mathbb{R}^{R \times \Delta T \times W}$.

4.3.2 Network architecture

Node embedder module: Receiving as input the initial feature tensor $\mathbf{F}$, this module computes latent contextual representations of dynamic FC features (Fig. 4.2). First, dynamic FC features between pairs of ROIs are extracted as: $\mathbf{dFC}_{r_i, r_j, w} = \text{Corr}(\mathbf{F}_{r_i, \cdot, w}, \mathbf{F}_{r_j, \cdot, w})$ where $w \in \{1, \dots, W\}$ denotes window index, $r_i \in \{1, \dots, R\}$, $r_j \in \{1, \dots, R\}$ denote ROI indices. The extracted features are then processed with a transformer encoder that learns contextual representations via attention between tokens in each window. Here, each token is taken as an individual ROI, and the input vector for each token is taken as its dynamic FC features with remaining ROIs (i.e., $\mathbf{dFC}_{r_i, \cdot, w}$ for $\text{ROI}(r_i)$). To enable attention calculations, the dynamic FC features concatenated across tokens (i.e., ROIs) are subjected to layer normalization (LN), and then window-specific matrices for keys $K_w \in \mathbb{R}^{R \times d}$, queries $Q_w \in \mathbb{R}^{R \times d}$ and values $V_w \in \mathbb{R}^{R \times d}$ are derived with learnable linear projection matrices $U_q, U_k, U_v \in \mathbb{R}^{R \times d}$:

$$\begin{aligned} Q_w &= \text{LN}([\mathbf{dFC}_{r_1, \cdot, w}, \mathbf{dFC}_{r_2, \cdot, w}, \dots, \mathbf{dFC}_{r_R, \cdot, w}]) U_q, \\ K_w &= \text{LN}([\mathbf{dFC}_{r_1, \cdot, w}, \mathbf{dFC}_{r_2, \cdot, w}, \dots, \mathbf{dFC}_{r_R, \cdot, w}]) U_k, \\ V_w &= \text{LN}([\mathbf{dFC}_{r_1, \cdot, w}, \mathbf{dFC}_{r_2, \cdot, w}, \dots, \mathbf{dFC}_{r_R, \cdot, w}]) U_v, \end{aligned} \quad (4.2)$$

where d is the embedding dimensionality for tokens. Note that the above computations are performed separately for each window. The window-specific attention matrix $\mathbf{A}_w \in \mathbb{R}^{R \times R}$ can then be computed as [233]:

$$\mathbf{A}_w = \text{Att}(Q_w, K_w, V_w) = \text{Softmax}\left(\frac{Q_w K_w^\top}{\sqrt{d}}\right) V_w, \quad (4.3)$$

where $^\top$ denotes matrix transpose. Note that we do not use any position encoding for the tokens since we observed in early phases of the study that inclusion of position encoding does not yield notable performance differences. Next, the

window-specific attention matrix is propagated to an MLP block following normalization:

$$\mathbf{EMB}_{\cdot, \cdot, w} = \text{MLP}(\text{LN}(\mathbf{A}_w)) = \text{GELU}(\text{LN}(\mathbf{A}_w)\mathbf{M}_1)\mathbf{M}_2, \quad (4.4)$$

where $\mathbf{M}_1 \in \mathbb{R}^{R \times D}$ and $\mathbf{M}_2 \in \mathbb{R}^{D \times D}$ denote MLP model parameters and GELU is a Gaussian activation function. In Eq. 4.4, $\mathbf{EMB} \in \mathbb{R}^{R \times D \times W}$ denotes the output embedding tensor with D taken as embedding dimensionality such that $D < R$.

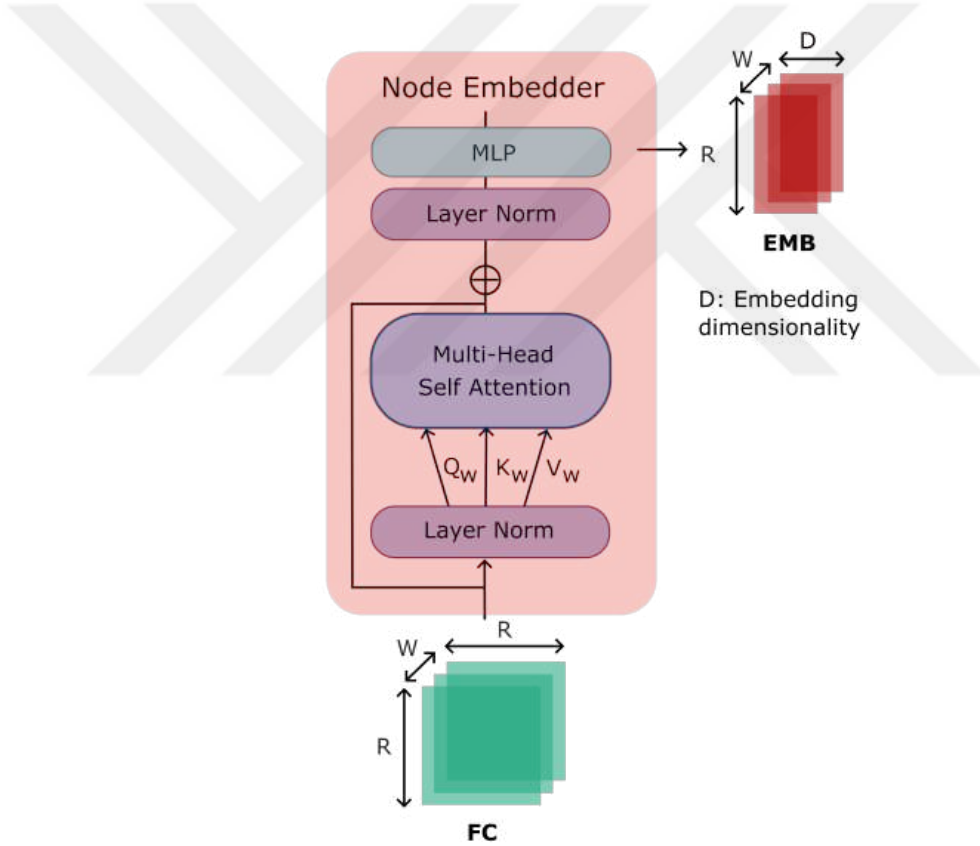


Figure 4.2: The node embedder module. The module input is a tensor of time-windowed dynamic FC features $\mathbf{dFC} \in \mathbb{R}^{R \times R \times W}$, with R denoting the number of ROIs, W denoting the number of time windows. The input tensor is processed via a transformer encoder with multi-head self-attention (MHSA), layer normalization (LN), and multi-layer perceptron (MLP) layers. Processing is performed for each time window separately. The module output is a node embedding tensor $\mathbf{EMB} \in \mathbb{R}^{R \times D \times W}$ where D is embedding dimensionality.

Lag filter module: Receiving as input the tensor of node features $\mathbf{F}$ (i.e., tensor of time-windowed BOLD signals), this module computes delayed connectivity features between pairs of ROIs across a range of temporal lags (Fig. 4.3).

To store delayed versions of signals within each window, the input tensor is first zero-padded along the second dimension that spans across the time frames within each window:

$$\mathbf{X} = [\mathbf{0}_{(R \times m \times W)}, \mathbf{F}_{(R \times \Delta T \times W)}, \mathbf{0}_{(R \times m \times W)}], \quad (4.5)$$

where $\mathbf{X} \in \mathbb{R}^{R \times (\Delta T + 2m) \times W}$, and m is set to define the range of time delays $\tau \in \{-m, -m+1, \dots, m-1, m\}$ that will be considered in the module. Given a pair of nodes, the window-specific cross-correlation between their BOLD signals is computed at each lag value separately:

$$\rho_{r_i, r_j, w, \tau} = (\mathbf{X}_{r_i, \cdot, w} \star \mathbf{X}_{r_j, \cdot, w})[\tau], \quad (4.6)$$

where $\star$ denotes the cross-correlation operator, and $\rho \in \mathbb{R}^{R \times R \times W \times (2m+1)}$ is the cross-correlation tensor. Afterwards, the cross-correlation vectors between each pair of connected nodes (i.e., $\langle \text{ROI}(r_i), \text{ROI}(r_j) \rangle \in E$) and each window are mapped onto lag activations. This mapping is performed via learnable lag filters $\mathbf{M}_{LF} \in \mathbb{R}^{(2m+1) \times k}$ with k denoting the number of filters:

$$\mathbf{LAG}_{r_i, r_j, w, \cdot} = \text{GELU}(\rho_{r_i, r_j, w, \cdot} \mathbf{M}_{LF}) \quad (4.7)$$

where $\mathbf{LAG} \in \mathbb{R}^{R \times R \times W \times k}$ is the lag activation tensor. Note that $\mathbf{LAG}_{r_i, r_j, \cdot, \cdot}$ is taken as 0 for $\langle \text{ROI}(r_i), \text{ROI}(r_j) \rangle \notin E$.

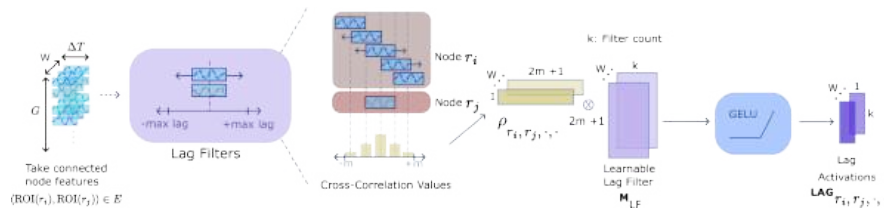


Figure 4.3: The lag filter module. The module input is a tensor of time-windowed BOLD signals $\mathbf{F} \in \mathbb{R}^{R \times \Delta T \times W}$, with ΔT denoting the duration of each window. Delayed connectivity features of pairs of ROIs are computed as the cross-correlation of their BOLD signals, with delays $\tau \in \{-m, -m+1, \dots, m-1, m\}$. Afterwards, the cross-correlation tensor $\rho \in \mathbb{R}^{R \times R \times W \times (2m+1)}$ is linearly transformed with learnable lag filters $\mathbf{M}_{LF} \in \mathbb{R}^{(2m+1) \times k}$ where k denotes the number of filters. The module output is the lag activation tensor $\mathbf{LAG} \in \mathbb{R}^{R \times R \times W \times k}$.

4.3.3 Graph learning

The node embedder module produces an embedding tensor, **EMB**, that reflects instantaneous inter-regional correlations between graph nodes, whereas the lag filter produces a lag activation tensor, **LAG**, that reflects delayed inter-regional correlations between nodes. To consolidate these two feature sets on the graph, node embeddings are taken as node features and lag activations are taken as edge weights (Fig. 4.1). To compute enhanced features, a message passing algorithm is then run on the graph to fuse the feature sets [221]. For this purpose, a four-dimensional message tensor $\mathbf{MES} \in \mathbb{R}^{R \times R \times (D \times k) \times W}$ is derived. Window-specific messages aimed at a target r_i -th node originating from its neighbors $r_j \in \mathcal{N}(r_i)$ are computed as:

$$\mathbf{MES}_{r_i, r_j, \cdot, w} = \text{flatten}(\mathbf{EMB}_{r_i, \cdot, w} \otimes \mathbf{LAG}_{r_i, r_j, w, \cdot}), \quad (4.8)$$

with $\otimes$ denoting outer product, and $\text{flatten}(\cdot)$ refers to vectorization. The incoming messages at the target node are averaged across both windows and neighbours:

$$\mathbf{AGG}_{r_i, \cdot} = \sum_{r_j \in \mathcal{N}(r_i)} \frac{1}{W} \sum_{w=1}^W \mathbf{MES}_{r_i, r_j, \cdot, w}, \quad (4.9)$$

where $\mathbf{AGG} \in \mathbb{R}^{R \times (D \times k)}$ is the aggregate message matrix for the graph. The aggregate message is concatenated with the window-averaged node embedding at each node $\text{ROI}(r_i)$:

$$\mathbf{OUT}_{r_i, \cdot} = \left[\frac{1}{W} \sum_{w=1}^W \mathbf{EMB}_{r_i, \cdot, w}, \mathbf{AGG}_{r_i, \cdot} \right]. \quad (4.10)$$

As such, $\mathbf{OUT} \in \mathbb{R}^{R \times (Dk + D)}$ denotes the enhanced feature matrix for the graph fusing features from node embedder and lag filter modules.

4.4 Experiments

4.4.1 Experimental procedures

Demonstrations were performed on fMRI data from the HCP S1200 release⁴ [103] and ID1000 dataset from Amsterdam Open MRI Collection (AOMIC)⁵ [210]. From the HCP S1200 release, preprocessed data from resting-state fMRI scans (HCP-Rest) and from task-based fMRI scans (HCP-Task) were analyzed. In HCP-Rest, the first resting-state scan among four sessions was selected for each subject, excluding short scans with $T < 1200$. This resulted in a total of 1093 healthy subjects (594 female and 499 male). In HCP-Task, scans recorded while participants performed 7 different cognitive tasks (emotion, relational, gambling, language, social, motor, working memory) were analyzed, resulting in a total of 1095 healthy subjects (594 female and 501 male). In the ID1000 dataset, preprocessed data from fMRI scans recorded during movie watching were analyzed. All scans had a fixed duration of $T = 240$. A total of 881 healthy subjects were examined (458 female and 423 male). For all datasets, two alternative sets of ROI definitions were obtained based on commonly adopted brain atlases in the neuroimaging literature. In particular, we considered the Schaefer atlas that provides functional-connectivity-based definitions of $R = 400$ ROIs [150], and the AAL atlas that provides anatomy-based definitions of $R = 116$ ROIs [151]. Since these atlases delineated ROI boundaries based on anatomical versus functional criteria and included different numbers of ROIs, they allowed us to examine performance under diverse settings.

Experiments were conducted on a single NVIDIA Titan Xp GPU using the PyTorch framework. A nested cross-validation procedure was performed with 5 outer and 1 inner folds. Data were three-way split into a training set (70%), a validation set (10%) and a test set (20%) with no subject overlap between the sets. For fair comparison, all models were trained, validated and tested on identical data splits. All models were trained based on cross-entropy loss. For each model,

⁴<https://db.humanconnectome.org>

⁵<https://openneuro.org/datasets/ds003097/versions/1.2.1>

hyperparameters were selected to maximize the average performance across the validation sets [126]. A common set of hyperparameters that were observed to yield near-optimal performance were used across datasets and atlases.

4.4.2 Comparison studies

GraphCorr was comparatively demonstrated against alternative plug-in methods for state-of-the-art baseline classification models. In all graph models, ROIs in a given brain atlas were taken as nodes, and edge selection was then performed based on Pearson’s correlation coefficient between BOLD signals. Edges whose correlation coefficients were in the top $z = 2\%$ were retained, while remaining edges were discarded [126]. Hyperparameters of both baseline models and plug-in methods were selected to optimize validation performance. Implementation details are given below.

4.4.2.1 Baseline models

Several graphical and convolutional baseline models were examined for gender detection and cognitive-task detection from fMRI scans.

SAGE. SAGE is a GNN model based on modules containing graph convolution, pooling and fully-connected layers [211]. A cascade of two graphical modules was used with a hidden dimension of 250 and a dropout rate of 0.5. Cross-validated hyperparameters were a learning rate of 3e-3, 20 epochs, and a batch size of 12.

GCN. GCN is a GNN model based on modules containing graph convolution, pooling and fully-connected layers [197]. A cascade of two graphical modules was used with a hidden dimension of 100 and a dropout rate of 0.5. Cross-validated hyperparameters were a learning rate of 5e-3, 30 epochs, and a batch size of 12.

BrainNetCNN. BrainNetCNN is a CNN model based on convolutional layers

with edge-to-edge and edge-to-node filters [91]. A hidden dimension of 32 and a dropout rate of 0.1 were used. Cross-validated hyperparameters were a learning rate of 2e-4, 20 epochs, and a batch size of 16.

4.4.2.2 Plug-in methods

Competing plug-in methods were used to augment baseline models with different input features. Identical modeling procedures were used for vanilla and augmented variants.

Vanilla. A vanilla variant was considered based on static FC features $\mathbf{sFC} \in \mathbb{R}^{R \times R}$ as model input.

Dynamic FC. An augmented variant was considered that computed dynamic FC features across separate time windows $\mathbf{dFC} \in \mathbb{R}^{R \times R \times W}$ via conventional correlation measures [75]. Time window definitions were matched to those in GraphCorr.

GRU. An augmented variant was considered that computed temporally-enhanced FC features via an RNN model based on GRU layers [234]. The output features had matching dimensionality to that of GraphCorr.

GraphCorr. The node embedder module was built with a single-layer transformer encoder. Because scan durations differed across datasets, dataset-specific ΔT (window size) and s (stride) were selected while common D (dimensionality), m (maximum lag) and k (filter count) were used. Note that $(Dk + D) \sim R$ was prescribed to avoid increases in dimensionality over static FC features. ($\Delta T=50$, $s=30$, $D=50$, $m=5$, $k=3$) were used for HCP-Rest and HCP-Task, while ($\Delta T=40$, $s=15$, $D=50$, $m=5$, $k=3$) were used for ID1000.

4.4.3 Explanatory analysis

To assess the influence of GraphCorr on interpretability, vanilla and augmented variants of baseline classification models were examined via an explanation procedure to identify the functional connectivity features that most saliently contribute to the model decisions. For this purpose, a gradient-based approach was employed to extract saliency scores for each model summarizing its important input features [235, 204, 236].

For the vanilla variant, gradients were computed with respect to the input static FC features **sFC**:

$$\mathbf{SAL}_{r_i, r_j}^{van} = |\nabla_{\mathbf{sFC}_{r_i, r_j}} y|, \quad (4.11)$$

where $\mathbf{SAL}^{van} \in \mathbb{R}^{R \times R}$ is the gradient matrix across inter-regional connections, and y denotes the model prediction. $\mathbf{SAL}^{van}$ was averaged across the column dimension to obtain a gradient vector across ROIs:

$$\mathbf{rSAL}^{van} = \sum_{r_j=1}^R \mathbf{SAL}_{\cdot, r_j}^{van}, \quad (4.12)$$

where $\mathbf{rSAL}^{van} \in \mathbb{R}^R$. For the augmented variants, gradients were computed with respect to the dynamic FC features **dFC**:

$$\mathbf{SAL}_{r_i, r_j, w}^{aug} = |\nabla_{\mathbf{dFC}_{r_i, r_j, w}} y|, \quad (4.13)$$

where $\mathbf{SAL}^{aug} \in \mathbb{R}^{R \times R \times W}$ is the gradient tensor across window-specific inter-regional connections. $\mathbf{SAL}^{aug}$ was averaged across the ROI (2nd) and window (3rd) dimensions to obtain a gradient vector $\mathbf{rSAL}^{aug} \in \mathbb{R}^R$ across ROIs:

$$\mathbf{rSAL}^{aug} = \sum_{r_j=1}^R \left(\frac{1}{W} \sum_{w=1}^W \mathbf{SAL}_{\cdot, r_j, w}^{aug} \right) \quad (4.14)$$

For reliable inference, significant ROIs with absolute gradient values greater than zero were determined via a Wilcoxon signed-rank test across the test set [237]. ROI-specific saliency scores were taken as the negative log of p-values, and normalized to yield a summed score of 1 across ROIs.

Next, we assessed the consistency between the cortical distributions of ROI-specific saliency scores for a particular classification task (e.g., gender detection from fMRI scans), and ROI-specific importance scores for related cognitive variables (e.g., female, male) given neuroimaging literature. For a systematic assessment, we employed the NeuroSynth framework devised for meta-analysis of existing neuroimaging findings [238]. This framework analyzes the coordinates of activation within articles containing query cognitive variables, and returns importance scores that reflect the representation strength of those variables across ROIs. We examined the similarity between cortical maps of saliency scores and importance scores. This analysis was performed on ROIs with importance scores in the top 15%.

4.5 Results

4.5.1 Comparison studies

GraphCorr was demonstrated on several state-of-the-art baseline classification models for fMRI analysis including SAGE [211], GCN [197], and BrainNetCNN [91]. Vanilla variants of baseline models based on static FC features were compared against augmented variants based on dynamic FC features, GRU and GraphCorr (see Methods). Performances of vanilla and augmented variants of baseline models for gender detection on HCP-Rest and ID1000, and for cognitive-task detection on HCP-Task datasets are listed in Table 4.1 for the Schaefer atlas, and in Table 4.2 for the AAL atlas. When using the Schaefer atlas, GraphCorr outperforms competing methods in all cases ($p < 0.05$, Wilcoxon signed-rank test), except for HCP-Task where all methods generally yield similar ROC saturated near 100%. On average across baseline models and datasets, GraphCorr enables (accuracy, ROC)% improvements of (8.6, 6.8)% over the vanilla variant, (7.4, 5.2)% over the dynamic FC plug-in, and (9.3, 7.3)% over the GRU plug-in. When using the AAL atlas, GraphCorr again outperforms competing plug-ins in all cases ($p < 0.05$), except for GRU that yields occasionally yields similar ROC on

HCP-Task. On average, GraphCorr enables improvements of (13.8, 10.0)% over the vanilla variant, (9.6, 7.2)% over the dynamic FC plug-in, and (9.4, 6.3)% over the GRU plug-in. Taken together, these results suggest that GraphCorr captures an enhanced set of FC features to augment baseline models for fMRI analysis, and thereby it improves classification performance over competing plug-in methods.

We also observe that vanilla variants of the relatively simpler graphical models (SAGE and GCN) perform poorly against the more complex BrainNetCNN model. However, GraphCorr-augmented variants of these graphical models can perform competitively with the augmented BrainNetCNN. For the Schaefer atlas, the average performance gap between BrainNetCNN and graphical models is lowered from (5.3, 6.3)% in the vanilla variants to (-0.1, 1.0)% in the GraphCorr-augmented variants. For the AAL atlas, the average performance gap between BrainNetCNN and graphical models is lowered from (5.1, 4.2)% in the vanilla variants to (-0.3, 0.7)% in the GraphCorr-augmented variants. These results suggest that the feature extraction capabilities of vanilla GNN models might be suboptimal in comparison to CNN models, albeit a powerful plug-in on the input side can mitigate this deficit.

Table 4.1: Performance of baseline models for gender detection on HCP-Rest and ID1000, and for cognitive task classification on HCP-Task based on the Schaefer atlas. Accuracy and ROC are listed as mean $\pm$ std across test folds for non-augmented (Vanilla), dynamic FC augmented (Dyn. FC), GRU augmented (GRU), and GraphCorr augmented (GraphCorr) variants. Boldface indicates the top performing plug-in for each model.

Model	Plug-in	HCP-Rest		ID1000		HCP-Task	
		Acc (%)	ROC (%)	Acc (%)	ROC (%)	Acc (%)	ROC (%)
SAGE	Vanilla	75.2 $\pm$ 2.8	85.3 $\pm$ 1.7	62.4 $\pm$ 2.2	68.6 $\pm$ 3.8	97.1 $\pm$ 0.4	99.9$\pm$0.0
	Dyn. FC	78.0 $\pm$ 2.2	86.6 $\pm$ 1.7	67.4 $\pm$ 1.3	73.5 $\pm$ 2.6	95.3 $\pm$ 1.0	99.7 $\pm$ 0.1
	GRU	79.2 $\pm$ 2.9	87.6 $\pm$ 2.6	65.5 $\pm$ 1.4	70.2 $\pm$ 1.3	97.8 $\pm$ 0.4	99.7 $\pm$ 0.1
	GraphCorr	89.6$\pm$0.7	94.3$\pm$2.0	81.7$\pm$2.7	87.0$\pm$2.1	99.2$\pm$0.2	99.5 $\pm$ 0.2
GCN	Vanilla	79.1 $\pm$ 2.9	86.0 $\pm$ 1.5	67.8 $\pm$ 3.0	71.8 $\pm$ 3.9	96.7 $\pm$ 0.5	99.8$\pm$0.0
	Dyn. FC	79.6 $\pm$ 2.6	88.0 $\pm$ 2.7	67.6 $\pm$ 4.0	74.5 $\pm$ 3.4	94.8 $\pm$ 0.8	99.7 $\pm$ 0.1
	GRU	78.3 $\pm$ 2.0	85.6 $\pm$ 1.1	62.5 $\pm$ 0.7	67.9 $\pm$ 2.2	97.5 $\pm$ 0.5	99.8$\pm$0.1
	GraphCorr	89.9$\pm$2.2	94.5$\pm$1.6	80.8$\pm$1.0	87.9$\pm$1.8	99.3$\pm$0.2	99.8$\pm$0.1
BrainNetCNN	Vanilla	82.5 $\pm$ 2.8	91.2 $\pm$ 1.2	75.5 $\pm$ 2.0	83.7 $\pm$ 2.1	97.0 $\pm$ 0.2	99.8 $\pm$ 0.0
	Dyn. FC	84.5 $\pm$ 2.9	91.7 $\pm$ 1.1	79.8 $\pm$ 2.1	87.1 $\pm$ 1.4	96.6 $\pm$ 0.7	99.8 $\pm$ 0.1
	GRU	78.7 $\pm$ 2.8	87.4 $\pm$ 4.2	76.3 $\pm$ 1.2	84.1 $\pm$ 1.5	90.8 $\pm$ 1.9	99.3 $\pm$ 0.2
	GraphCorr	88.5$\pm$2.6	94.7$\pm$1.9	82.7$\pm$1.6	89.9$\pm$2.2	98.8$\pm$0.2	99.9$\pm$0.0

Table 4.2: Performance of baseline models for gender detection on HCP-Rest and ID1000, and for cognitive task classification on HCP-Task based on the AAL atlas.

Model	Plug-in	HCP-Rest		ID1000		HCP-Task	
		Acc (%)	ROC (%)	Acc (%)	ROC (%)	Acc (%)	ROC (%)
SAGE	Vanilla	68.3±3.3	75.7±1.3	62.8±1.8	67.2±2.7	81.0±1.1	97.2±0.1
	Dyn. FC	69.0±3.2	76.4±2.3	67.1±4.0	72.6±2.9	87.3±0.9	98.6±0.2
	GRU	76.2±2.3	82.7±1.4	68.6±3.3	73.7±2.0	95.3±1.4	99.6±0.2
	GraphCorr	85.5±3.6	91.2±2.6	77.5±3.7	84.9±1.9	95.8±0.4	99.6±0.1
GCN	Vanilla	69.9±1.4	75.9±1.1	65.5 ±1.4	71.0±1.2	77.5±1.0	95.9±0.2
	Dyn. FC	69.4±2.3	75.9±2.6	68.4±4.6	74.8±3.3	85.0±1.5	98.5±0.1
	GRU	74.5±2.9	82.8±1.9	67.2±2.2	72.0±2.7	94.9±0.6	99.7±0.0
	GraphCorr	84.4±3.2	89.0±2.8	79.4±3.5	85.4±2.5	96.1±0.4	99.9±0.1
BrainNetCNN	Vanilla	68.2±3.5	74.8±2.1	75.0±2.2	81.4±3.3	84.7±0.1	97.8±0.2
	Dyn. FC	76.7±1.7	83.6±1.0	76.3±3.1	83.3±2.4	91.3±0.5	99.1±0.1
	GRU	73.4±2.2	81.8±1.9	75.1±2.2	82.9±2.1	67.5±2.8	95.4±0.6
	GraphCorr	84.0±2.9	91.3±2.7	79.0±1.8	86.2±0.8	95.4±0.6	99.7±0.1

4.5.2 Explanatory analysis

To assess the influence of GraphCorr on interpretability, explanatory analyses were conducted on vanilla and augmented variants of baseline classification models. These analyses examined the importance of each input FC feature on the model decisions. For this purpose, the gradients of the model output with respect to individual FC features were computed and ROI-specific saliency scores were obtained from the gradients (see Methods). SAGE was selected as the baseline model since it generally maintains the highest performance after GraphCorr augmentation (Tables 4.1-4.2). Gender detection on HCP-Rest was analyzed that involves distinguishing between female and male classes. For cognitive-task detection on HCP-Task, the motor task was analyzed as a representative case that involves distinguishing between motor and remaining non-motor tasks. Reference

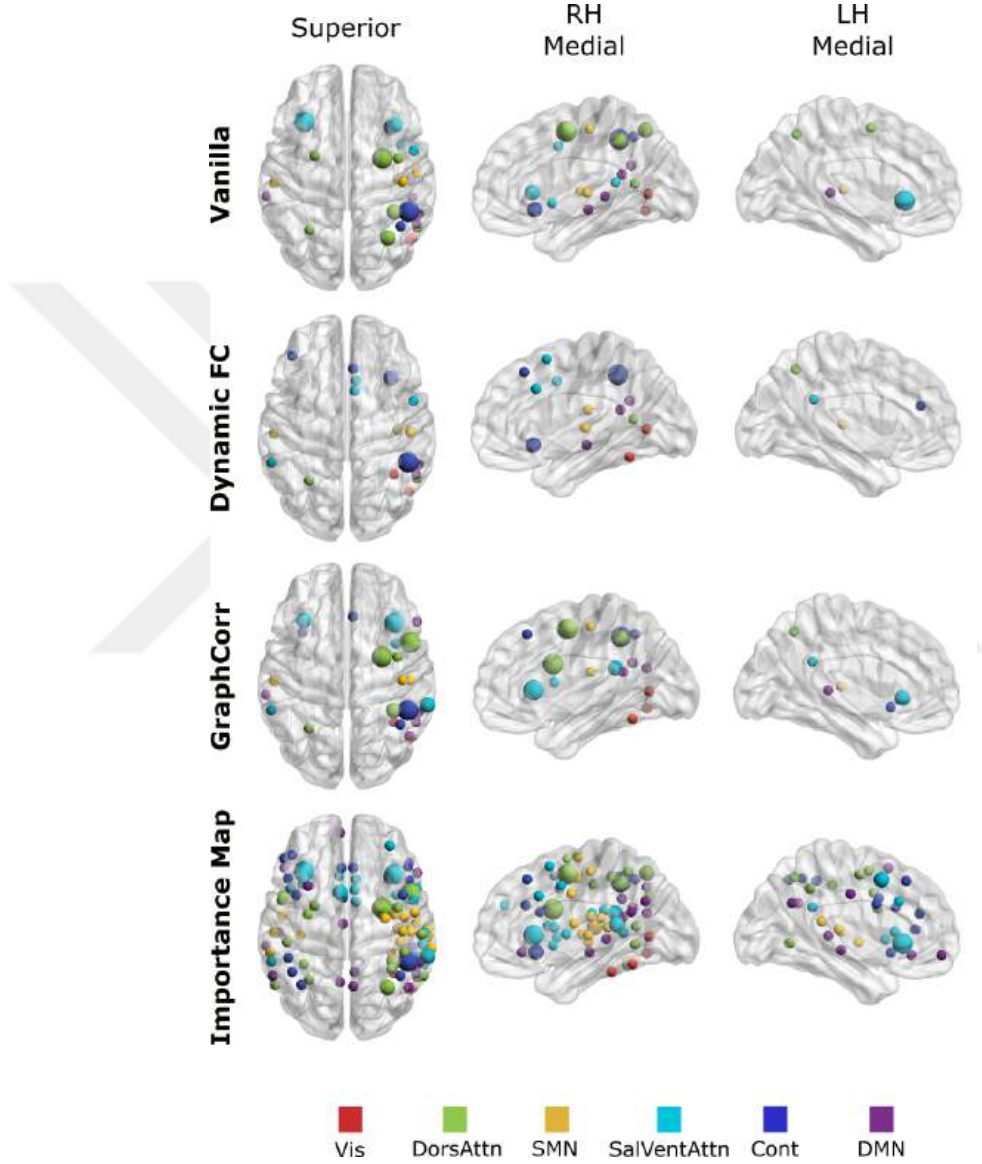


Figure 4.4: Explanatory analysis for gender detection classifiers. Saliency maps are shown for vanilla, dynamic FC- and GraphCorr-augmented SAGE models on the HCP-Rest dataset, along with reference importance maps derived via a meta-analysis of neuroimaging literature. Important ROIs are marked with colored spheres. In saliency maps, sphere size indicates the relative saliency score for each ROI. In importance maps, sphere size indicates the relative importance score. Color-coding indicates membership to functional networks: (Vis = Visual network, Attn = attention network, SMN = sensorimotor network, Cont = control network, DMN = default mode network.)

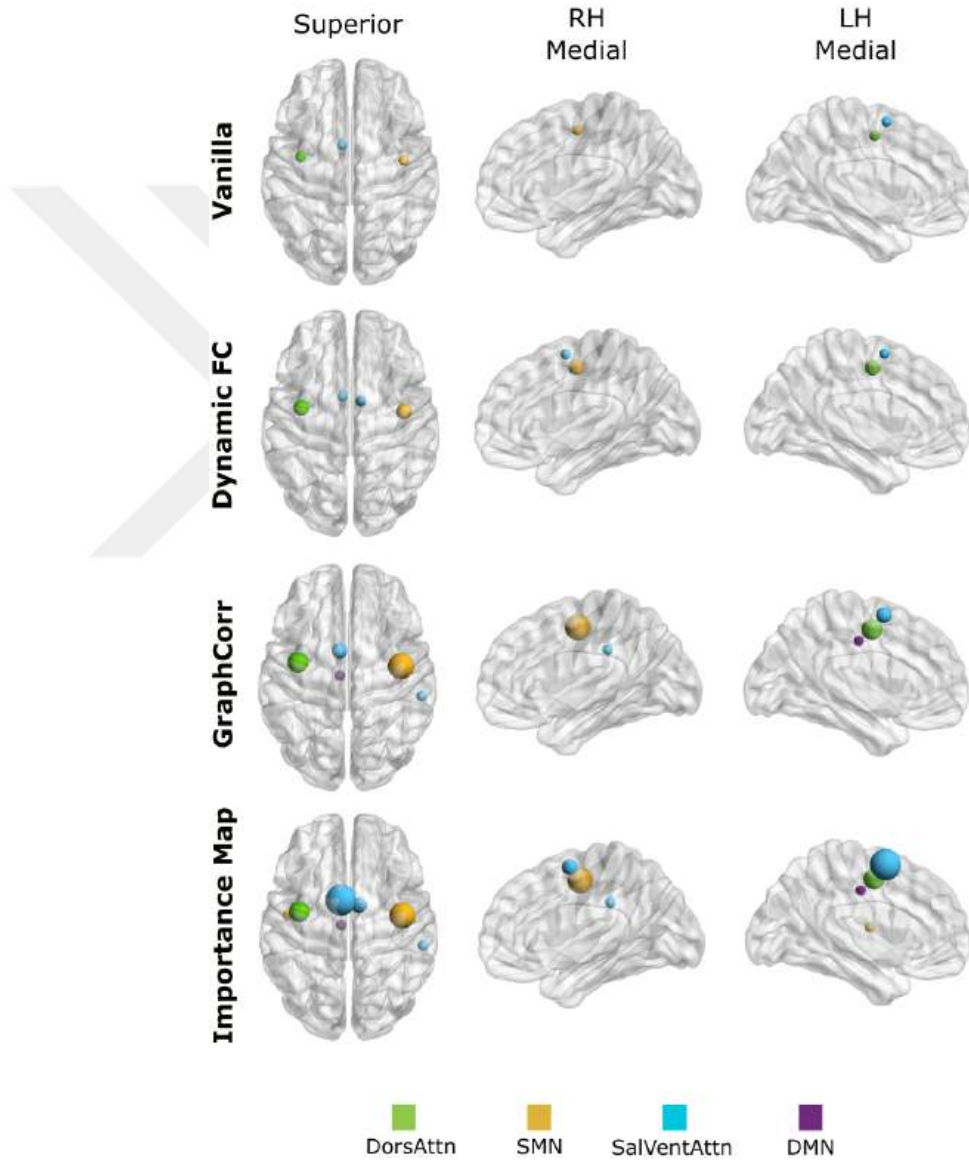


Figure 4.5: Explanatory analysis for cognitive-task detection classifiers. Saliency maps for the motor task are shown based on vanilla, dynamic FC- and GraphCorr-augmented SAGE models on the HCP-Task dataset, along with reference importance maps derived via a meta-analysis of neuroimaging literature.

importance scores of each ROI for representing variables related to the classification task (i.e., ‘female-male’ for gender detection, ‘motor task’ for cognitive-task detection) were derived from a meta-analysis via the NeuroSynth framework (see Methods). Cortical maps of saliency scores were visualized for the vanilla, dynamic FC-augmented and GraphCorr-augmented variants, along with reference cortical maps of importance scores.

Fig. 4.4 depicts results for gender detection. For the vanilla variant, saliency scores are significant in the left hemisphere (LH) for few regions within parietal, somatomotor, and prefrontal cortices; and for several regions within the right hemisphere (RH) for visual, parietal, somatomotor, temporal, and prefrontal cortices ($p < 0.05$, Wilcoxon signed-rank test). For the dynamic-FC augmented variant, saliency scores are significant in LH for few regions within parietal, somatomotor, and prefrontal cortices; and in RH for several regions within visual, parietal, somatomotor, temporal, and prefrontal cortices. Meanwhile, for the GraphCorr-augmented variant, saliency scores are significant in LH for a greater number of regions within prefrontal cortices, temporoparietal junction and frontal operculum insula; and in RH for regions within visual, ventral frontal, lateral parietal cortices, and frontal operculum insula. Compared to other variants, GraphCorr generally yields a more consistent cortical distribution of saliency scores with respect to the reference importance maps, especially near regions associated with attention, default mode and control networks. Literature suggests that these networks carry prominent information for discriminating between female and male subjects [209, 204, 239]. In particular, regions in the attention (frontal operculum insula) network are assumed to be involved in spatial processing, attentional control, and working memory. Meanwhile, regions in the default mode network (temporoparietal junction) are assumed to be involved in aspects of social cognition such as face recognition and emotion processing, and regions in the control network (prefrontal) are assumed to be involved in high-level cognition including decision making [156, 209, 236]. Previous studies have reported gender differences in these cognitive abilities, and in activations across associated regions [240, 209].

Fig. 4.5 depicts results for the motor task included in the cognitive-task detection analyses. Consistently across variants, saliency scores are generally significant in both hemispheres for few regions within precentral cortex, and in LH for few regions within supplementary motor cortex ($p < 0.05$, Wilcoxon signed-rank test). Yet, for the GraphCorr-augmented variant, saliency scores are also significant in LH for a region within cingulate cortex, and in RH for a region within supramarginal gyrus. Compared to other variants, GraphCorr generally yields a more consistent cortical distribution of saliency scores with respect to the reference importance maps, including regions associated with attention, sensorimotor, and default mode networks. Literature indicates that these networks carry prominent information for discriminating motor from non-motor cognitive tasks [209]. In particular, regions in the attention (precentral) network are assumed to be involved in motor planning, regions in the sensorimotor network (supplementary motor) are assumed to be involved in motor execution, and regions in the default mode network (cingulate, supramarginal) are assumed to be involved in motor control and association of motor and sensory data [204]. Taken together, these results suggest that GraphCorr elicits explanations that are more closely aligned with neuroscience findings in the literature.

4.5.3 Ablation studies

Ablation studies were performed to assess the contribution of the individual design elements in GraphCorr to model performance. These analyses were conducted based on the SAGE model using the HCP-Rest dataset and the Schaefer atlas, i.e., the setting that yields the highest overall performance for gender detection. First, we assessed contributions of the node embedder module, lag filter module, and time windowing in GraphCorr. To ablate the node embedder module, node embeddings prior to message passing were initialized with the unlearned time-windowed FC matrix derived via conventional correlation measures on BOLD signals. To ablate the lag filter module, a single filter at zero lag was used within the module to consider only instantaneous correlations. To ablate time windowing, the entire fMRI time series was provided to GraphCorr with a

single window of size equal to the scan duration (i.e., $\Delta T=1200$). Table 4.3 lists performance metrics for ablated variants of GraphCorr. We find that the node embedder module, the lag filter module and time windowing enable (accuracy, ROC)% improvements of (5.3, 3.2)%, (0.7, 0.2)%, and (8.4, 5.6)%, respectively. These results suggests that all design elements contribute to improving model performance.

Next, we evaluated GraphCorr variants obtained by employing different time windowing procedures on the original fMRI time series. In particular, we examined the influence of the window size (ΔT) and stride (s) parameters on gender detection performance. Separate variants were trained for $\Delta T=[25\ 1000]$ while $s=30$, and for $s=[25\ 50]$ while $\Delta T=50$. Table 4.4 lists performance metrics for GraphCorr variants based on the SAGE model using the HCP-Rest dataset and the Schaefer atlas. We find that the values of window size and stride selected based on validation performance yield near-optimal performance on the test set, and that there are modest performance variations with changing window size or stride in the immediate neighborhood of selected values. That said, for window size, a notable performance drop occurs as early as $\Delta T=200$ and grows further towards larger ΔT . For stride, a notable performance drop occurs at $s=50$.

Table 4.3: Performance for GraphCorr variants ablated of individual design elements in gender detection. Results listed for the augmented SAGE model on the HCP-Rest dataset with the Schaefer atlas.

Node Embed- der	Lag Filter	Windowing	Acc (%)	ROC (%)
X	X	X	75.2 $\pm$ 2.8	85.3 $\pm$ 1.7
✓	X	X	80.5 $\pm$ 1.9	88.5 $\pm$ 1.5
✓	✓	X	81.2 $\pm$ 1.7	88.7 $\pm$ 2.1
✓	✓	✓	89.6$\pm$0.7	94.3$\pm$2.0

Table 4.4: Performance for GraphCorr variants that use $\Delta T = [25\ 1000]$ while $s = 30$, and $s = [25\ 50]$ while $\Delta T = 50$ for gender detection. Results listed for the augmented SAGE model on HCP-Rest with the Schaefer atlas.

Parameter		Acc(%)	ROC(%)
ΔT ($s = 30$)	25	89.6$\pm$2.0	91.8 $\pm$ 3.6
	50	89.6$\pm$0.7	94.3$\pm$2.0
	75	89.6$\pm$2.2	92.0 $\pm$ 2.3
	100	89.4 $\pm$ 2.1	93.4 $\pm$ 0.9
	200	88.0 $\pm$ 2.3	94.0 $\pm$ 2.0
	600	85.9 $\pm$ 1.9	92.8 $\pm$ 0.7
	1000	82.3 $\pm$ 1.6	90.0 $\pm$ 1.5
s ($\Delta T = 50$)	25	89.5 $\pm$ 2.0	93.6 $\pm$ 2.1
	30	89.6$\pm$0.7	94.3$\pm$2.0
	40	88.9 $\pm$ 2.5	91.4 $\pm$ 3.0
	50	87.1 $\pm$ 4.5	90.1 $\pm$ 4.6

GraphCorr is devised to improve capture of information in time-varying FC patterns without expanding dimensionality compared to static FC (sFC) features. To assess whether GraphCorr-based features carry enhanced temporal information over sFC features, we conducted an analysis with conventional dynamic FC (dFC) features based on time-windowed correlation measures taken as reference. For improved fidelity in this analysis, the HCP-Task dataset was examined where BOLD signals follow a more structured time-course due to the task instructions given to subjects [103]. To obtain a basis for time-varying FC patterns in fMRI scans, we performed principal components analysis (PCA) on dFC features across time windows. The first 6 PCs that explain 95% of the variance in dFC features were selected. We reasoned that if a feature set carries enhanced temporal information, then it should have stronger projections onto the PCs of reference dFC features. Fig. 4.6 displays the PC projections for GraphCorr-based FC features and sFC features. We find that GraphCorr generally elicits higher projections onto each individual PC, corroborating that it offers enhanced capture of time-varying FC patterns in its output features.

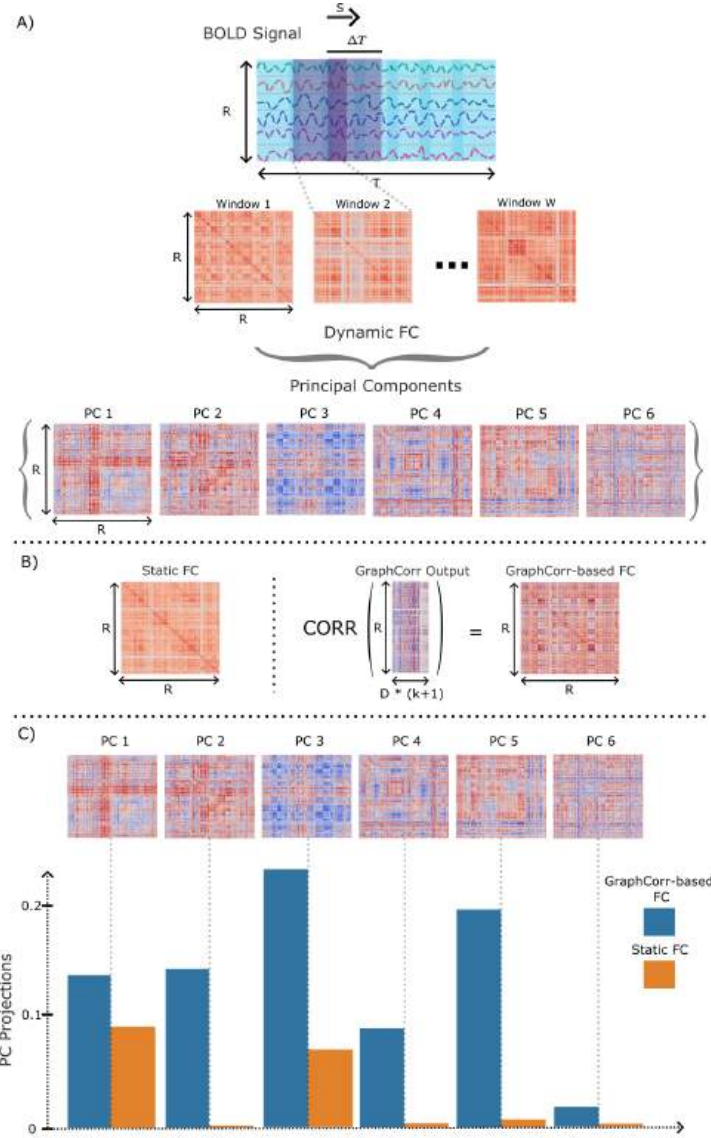


Figure 4.6: Analysis of temporal sensitivity for GraphCorr on HCP-Task with the Schaefer atlas. A) As reference for time-varying FC patterns in fMRI scans, conventional dynamic FC (dFC) features were computed via time-windowed correlation measures on BOLD signals. Principal components analysis (PCA) was performed on the dFC features across time windows to obtain a basis for time-varying FC patterns. The first 6 PCs that explain 95% of the variance were selected. B) Static FC (sFC) features and GraphCorr-based FC features were extracted. Since GraphCorr outputs features in a latent embedding space, FC features were derived via correlation measures between the embedding vectors for pairs of ROIs. C) GraphCorr-based FC and sFC features were projected onto the PCs. In general, GraphCorr yields higher projections onto the PCs than sFC.

4.6 Discussion

Here we reported a novel plug-in GNN method, GraphCorr, to improve the performance of classification models for fMRI analysis by capturing dynamic, lagged FC features of BOLD signals. Demonstrations were provided on three large-scale fMRI datasets, where substantially improved performance was achieved following model augmentation with GraphCorr. For explanation of these results, the cortical distribution of saliency scores for each augmentation method was compared against NeuroSynth-based reference importance maps derived from a diverse collection of studies that use task-based and resting-state fMRI as well as other modalities. While a native degree of discrepancy can be present with the datasets examined here, NeuroSynth enables large-scale analyses with statistical power exceeding that of isolated analyses on few tens of subjects in individual studies [241], so it serves as a reliable benchmark to interpret neuroimaging findings [242, 243, 244]. Compared to vanilla and dynamic-FC augmented variants of classification models, we find that the GraphCorr-augmented variant yields a greater number of significant regions whose saliency scores are closely aligned with reference importance maps. Note that the saliency maps are derived based on model gradients with respect to input features, so both the level of task-relevant information in input features and the learning accuracy for model weights can influence the significance of brain regions. For the vanilla variant, the limited number of significant regions can be attributed to the lack of temporal information in sFC features on dynamic changes across fMRI scans. While the dynamic FC method provides enhanced temporal information to the classification model, the high dimensionality of dFC features can compromise accuracy of learned model weights and thereby gradients. Thus, the relatively lower number of significant regions might be attributed to these inaccuracies.

A mainstream approach in neuroimaging studies rests on prediction of experimental variables typically related to stimulus or task from BOLD signals [107, 43]. Here we adopted this approach to build decoding models that predict subject gender and cognitive task from fMRI scans. The proposed method can also be combined with classification models to predict other categorical variables

related to disease [126, 91] or continuous variables related to stimulus or task features [45]. An alternative procedure to examine cortical function rests on encoding models that instead predict BOLD signals from experimental variables [50, 48, 167, 245]. It may be possible to adopt GraphCorr to improve sensitivity of such baseline encoding models. In this case, GraphCorr would receive as input the time course of experimental variables during an fMRI scan. In turn, it would learn dynamic, lagged correlations among experimental variables to better account for their time-varying distribution. Learned correlations might help improve performance of regression models that aim to predict measured BOLD signals. Future work is warranted to investigate the potential of GraphCorr in building encoding models for fMRI.

An important parameter set for GraphCorr, related to its ability to enhance fine-grained temporal information in its output features, includes the window size ΔT and the window stride s . These parameters control the extent of time windows over which instantaneous interactions are computed in the node embedder module, and lagged interactions are computed in the lag filter module. In theory, lower ΔT and s should increase sensitivity to relatively rapid temporal changes in fMRI scans. In practice, however, the dynamic FC features input to GraphCorr are extracted via correlation measures on noisy measurements, so the accuracy of FC features can be degraded for shorter windows with fewer time frames. Given this intrinsic trade-off between temporal sensitivity and feature accuracy, different parameter values could be preferred based on the rate of temporal fluctuations and noise levels. Here we observed moderate differences in the optimal parameters for separate datasets, and a degree of reliability in classification performance under reasonable variations in parameter values. Yet, parameter tuning might serve a more critical role when analyzing datasets acquired at relatively low field strengths or high spatio-temporal resolutions.

Limitations

Several technical limitations related to GraphCorr can be addressed to further improve performance and efficiency in fMRI analysis. Here, GraphCorr was trained end-to-end with classification models on the HCP-Rest, ID1000, or HCP-Task datasets that each contain data from nearly a thousand subjects. While the lag filter module has low complexity, the node embedder module uses a transformer encoder with a relatively large number of parameters. In turn, high model complexity can elicit suboptimal learning on compact datasets of limited size. When learning deficiencies are suspected, transfer learning can be performed where the encoder is initialized with pre-trained weights [29]. Learning might also be enhanced via data augmentation procedures that can produce a large variety of realistic samples from a learned distribution [177, ?].

Another practical concern related to the transformer encoder in the node embedder module is the inference time, which scales quadratically with the length of the input sequence of tokens [98]. This quadratic scaling could yield notable burden while processing data from fMRI scans with significantly high spatial and/or temporal resolution. When computational burden becomes prohibitive, partitioning mechanisms on attention layers such as multi-query attention might be adopted to improve inference efficiency [246].

In this study, each subject’s fMRI scan was aligned to an anatomical template to define brain regions with guidance from an atlas. The mean BOLD signals across voxels in each ROI were then processed in plug-in methods and thereby classification models. Benefits of this approach include consistency in ROI definitions across subjects and computational efficiency due to relatively lower model complexity [173]. Yet, information losses naturally occur during registration of individual-subject fMRI data onto a standardized template. To alleviate these losses, ROI definitions in the template space could instead be backprojected onto the brain spaces of individual subjects. This alternative procedure can mediate ROI definitions while maintaining acquired fMRI data in its original anatomical space for improved spatial precision [43].

In GraphCorr, an initial graph is formulated where nodes are taken as ROIs defined according to a brain atlas, and binary edges are defined between ROIs whose static FC values exceed a certain threshold. For the fMRI datasets examined here, we observed that these binary edge definitions that are kept fixed during subsequent training procedures yield reasonably high performance. When desired, it may be possible to seek further performance improvements via an adaptive graph formulation where both the ROI definitions and the edge weights between the ROIs are taken as learnable parameters, at the expense of elevated model complexity.

Here, we primarily examined cross-subject generalization performance of vanilla and augmented classification models, where models were trained and tested on independent splits extracted from a given dataset. In certain applications, it may be desirable to transfer classification models across different imaging sites that may utilize different scanners and fMRI protocols. In such cases, data or model aggregation frameworks can be adopted to train models that reliably cope with shifts in the distribution of fMRI data across sites [219, 27, 247, 248].

Chapter 5

DreaMR: Diffusion-driven Counterfactual Explanation for Functional MRI

The content of this chapter reflects the study described in the following publications:

- **H. A. Bedel** and T. Çukur, "Dreamr: Diffusion-driven counterfactual explanation for functional MRI," arXiv preprint arXiv:2307.09547, 2023.
- **H. A. Bedel** and T. Çukur, "Explaining Deep fMRI Classifiers with Diffusion-Driven Counterfactual Generation," in *32st annual meeting of International Society for Magnetic Resonance Imaging (ISMRM)*, Toronto, May 2024.

5.1 Introduction

Functional magnetic resonance imaging (fMRI) enables non-invasive cognitive assessments by capturing time-varying blood-oxygen-level-dependent (BOLD) responses across the brain [249]. Spatiotemporally measured BOLD responses can be analyzed to infer associations between individual brain regions and cognition-related variables. A traditional framework in neuroscience performs inference via linear classifiers that are trained to predict variables given responses [65]. The classifier weight for a brain region is then taken to reflect the importance of that region in detecting the respective variable. Although this traditional approach offers ease of interpretation, linear classifiers typically suffer from poor sensitivity [82]. In recent years, deep-learning classifiers have gained prominence as they show substantially higher sensitivity to fine-grained patterns in fMRI data [250, 120, 91, 92, 90, 87, 128, 94, 251]. Despite this important benefit, hierarchical layers of nonlinear transformation in deep models obscure precise associations between brain responses and cognition-related variables, introducing an interpretation challenge and creating a barrier to methodological trust [252]. As such, there is a dire need for explanation methods that highlight the critically important set of input features (i.e., brain responses) for deep fMRI models to help interpret their decisions.

An emerging framework is counterfactual explanation that aims to identify a minimal, plausible set of changes in the features of an input fMRI sample (i.e., a subject’s fMRI scan) that is sufficient to alter the decision of a downstream analysis model [253]. To do this, a generative prior is commonly trained to capture the distribution of original fMRI samples such that new, random samples can be drawn from the learned data distribution [252]. Afterwards, the trained prior is employed to regenerate the values of spatiotemporal responses in an original input sample, in order to produce a counterfactual sample that is proximal to the original sample, albeit that changes the decision of the downstream model [254]. The difference between the two samples is then inspected to interpret associations between brain regions and cognitive state [253]. Counterfactual methods can offer superior feature specificity against attribution methods, which derive

gradient or activation heatmaps that can be broadly distributed across input features [204, 255]. They can also produce more sensitive interpretations against perturbation methods, which perform local degradations on input features that can disrupt global coherence [256, 257]. Nevertheless, the performance of counterfactual methods depend critically on the fidelity of samples synthesized by the underlying generative prior.

Previous studies on counterfactual explanation have commonly proposed variational autoencoder (VAE) or generative adversarial network (GAN) priors trained to capture the distribution of fMRI data [253, 254, 258]. These priors synthesize counterfactuals with high efficiency, albeit VAEs often suffer from relatively low sample quality due to loss of detailed features, and GANs suffer from training instabilities that can hamper sample quality or diversity [29]. A promising surrogate for sample generation is diffusion priors that offer high sample fidelity via a many-step sampling process [259, 260]. Few recent imaging studies have considered diffusion-based counterfactual generation to detect anomalous lesions in anatomical MRI and X-ray scans [261, 262, 263, 264]. Yet, to our knowledge, diffusion priors have not been explored for counterfactual explanation in multivariate fMRI analysis. This may be partly related to excessive inference times of conventional diffusion priors such as DDPM or DDIM that require hundreds of steps to generate a single sample [265]. When coupled with the high dimensionality of fMRI data, inefficiency of conventional diffusion priors can be particularly limiting in application domains that benefit from rapid data processing such as real-time fMRI or cohort fMRI studies [266, 267].

Here, we propose a novel diffusion-driven counterfactual explanation method, DreaMR, to interpret downstream fMRI classifiers with improved fidelity and efficiency. The proposed method trains a class-agnostic diffusion prior for fMRI data, and the trained prior generates a counterfactual sample with guidance from a downstream classifier to alter its decision (Fig. 5.1). To improve inference efficiency without compromising sample quality, DreaMR leverages a novel fractional multi-phase-distilled diffusion (FMD) prior that splits the diffusion process into consecutive fractions and performs multi-phase distillation in each fraction (Fig.

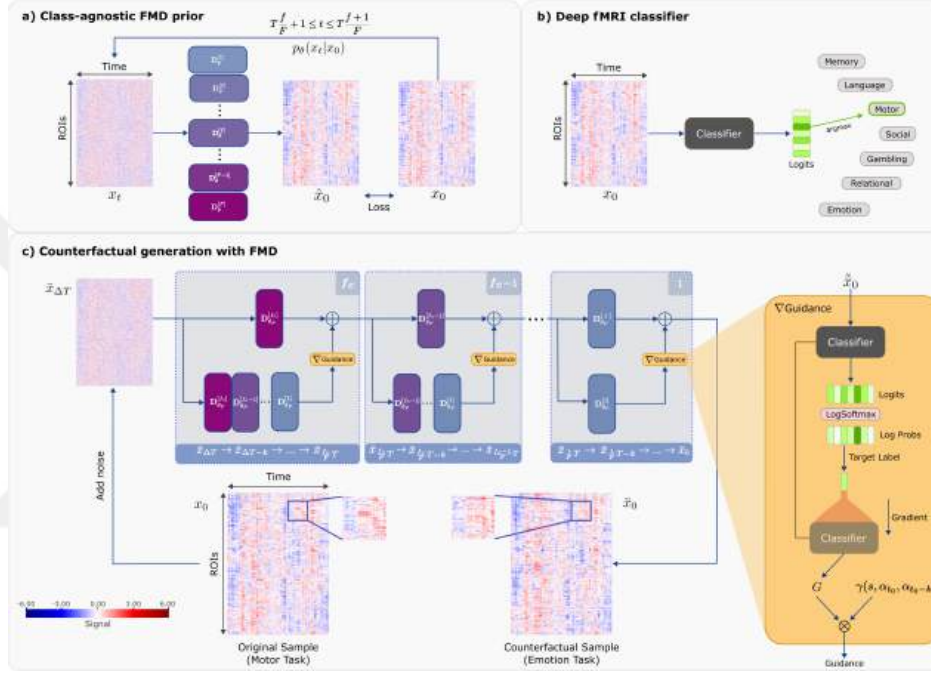


Figure 5.1: DreaMR is a counterfactual explanation method for deep fMRI classifiers. **a)** To capture the distribution of fMRI data, DreaMR trains a class-agnostic FMD prior that splits the diffusion process into F uniform fractions with dedicated networks $\mathbf{D}_\theta^{[f]}$, as described in Eq. 5.6. Following training, the FMD prior is subjected to multi-phase distillation to attain $\mathbf{D}_{\theta_P}^{[f]}$ that allow fast sampling, as described in Eq. 5.7. **b)** DreaMR is devised to explain the decisions of a deep classifier that predicts cognition-related variables from a subject’s fMRI scan. **c)** Given an input fMRI sample $x_0 \sim p(x)$ mapped onto cognitive state y_0 by the classifier, DreaMR first samples a noise-added version $\tilde{x}_{\Delta T}$ via forward diffusion and then generates $\tilde{x}_0 \sim p(x)$ with minimal alterations from x_0 via reverse diffusion. Generation is guided with the conditional score of the classifier to flip the decision to $\bar{y}_0 \neq y_0$. Starting at fraction $f_c = \lceil \Delta T(F/T) \rceil$, classifier guidance computed from a denoised estimate of the counterfactual sample ($\tilde{x}$) is injected at each diffusion step (orange boxes), as described in Alg. 2.

5.2). Unlike regular diffusion priors implemented with UNet backbone architectures, DreaMR uniquely leverages an efficient transformer architecture of linear complexity in conjunction with the FMD prior to capture long-range spatiotemporal context in fMRI scans. During counterfactual generation, classifier-guidance is injected separately into each diffusion fraction to tailor synthesis of intermediate samples in proximity of the original sample. The difference between the original and the final counterfactual samples reflects the contribution of brain regions to the model decision. We report comprehensive demonstrations to explain deep classifiers for sex detection on resting-state fMRI scans, and for cognitive task detection in task-based fMRI scans. We find that DreaMR achieves superior fidelity against competing explanation methods, and it substantially outperforms conventional diffusion priors in inference efficiency. Code to implement DreaMR is available at <https://github.com/icon-lab/DreaMR>.

Contributions

- To our knowledge, we introduce the first diffusion-driven counterfactual explanation method in the literature for multi-variate fMRI analysis.
- DreaMR generates counterfactual samples with a novel fractional multi-phase-distilled diffusion prior to boost inference efficiency without compromising quality.
- Unlike conventional diffusion methods, DreaMR leverages a transformer architecture of linear complexity to capture long-range spatiotemporal context in fMRI scans.

5.2 Related Work

5.2.1 Explanation of fMRI Models

Two prominent frameworks have been proposed in the literature to address the interpretation challenge for deep fMRI classifiers. A first framework is intrinsic interpretation where analyses are conducted using specialized models, such as linearized or graph classifiers, with restricted designs to permit an inherent degree of explanation [268, 94, 269]. Model-specific explanation is then attained by inspecting internal parameters, but use of restricted designs can elicit losses in classification performance [99, 200, 136, 1]. A more flexible framework is post-hoc explanation where analyses are conducted using an unrestricted model, and the influences of model inputs on the output are observed for interpretation. Among post-hoc techniques, attribution methods derive heatmaps across input features to estimate their salience, taken as gradients [270, 218], activations [271, 272, 273, 235], or a weighted combination of gradients and activations for the target class [204, 274, 255]. Attribution methods often require architecture-based modifications that limit practicality, and produce over-broad heatmaps that degrade interpretation specificity. To improve local specificity, perturbation methods introduce patch-level degradation on input features through operations such as occlusion [120, 94, 257, 256, 275, 276, 277]. Yet, perturbation-methods can be computationally heavy and they often produce globally-incoherent results that hamper the sensitivity of interpretations.

Counterfactual generation is an alternative post-hoc technique that resamples original data via a generative prior in order to enforce desired changes in input features [278, 279]. In the case of explaining fMRI classifiers, the aim of counterfactual generation would be to identify a minimally-sufficient set of changes in input features that flip the classifier decision [253]. Although improvements over attribution and perturbation methods have been reported for this application, VAE and GAN priors in previous counterfactual methods can suffer from low sample fidelity that in turn limits the reliability of explanations [252]. Recent machine learning studies advocate diffusion priors as a promising surrogate for

reliable sample generation, albeit conventional diffusion priors are known to suffer from an inherent trade-off between sampling quality and efficiency [259, 260]. This has impeded adoption of diffusion priors in counterfactual generation that requires iterated resampling of each high-dimensional fMRI data sample until the respective model decision is flipped. To address these open issues, here we introduce DreaMR, the first diffusion-driven explanation method for multivariate fMRI analysis to our knowledge.

5.2.2 Counterfactual Generation

Recent computer vision studies have considered diffusion priors to generate counterfactual natural images from desired object classes. [280, 281, 282] propose conventional diffusion priors based on a common UNet architecture. Interleaved sampling is used to trade-off sample quality in return for efficiency [281, 282]. Unlike conventional diffusion priors that typically require hundreds of diffusion steps even with interleaved sampling, DreaMR improves practicality via its novel FMD prior that achieves high sampling efficiency and quality via multi-phase distillation on consecutive fractions of the diffusion process, where each fraction uses a dedicated denoising network. To our knowledge, FMD is the first diffusion prior to perform fraction-specific distillation in the literature. Furthermore, DreaMR uniquely implements denoising networks via a transformer architecture to improve sensitivity to long-range context in fMRI scans that last several minutes [101].

Few recent imaging studies have also considered diffusion-based counterfactual generation to map lesions in anatomical scans by synthesizing pseudo-healthy medical images [261, 262, 263]. Commonly, these studies propose conventional diffusion priors based on UNet. Instead, DreaMR leverages the novel FMD prior for higher efficiency and a transformer architecture to capture long-range context. Note that [261, 262] train class-conditional priors on normals and patients, and [263] trains a class-specific prior on normals. When adopted for counterfactual explanation, such class-informed priors must use matching class definitions to

the classifier, so they require retraining for each classification task. In contrast, DreaMR employs a class-agnostic diffusion prior that can be utilized to explain models for different tasks without retraining.

5.3 Theory

5.3.1 Conventional Diffusion Priors

Diffusion priors use a gradual process to transform between a data sample x_0 and a random noise sample x_T in T steps. In the forward direction, Gaussian noise is added to obtain a noisier sample x_t at step t , with forward transition probability:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; (\alpha_t/\alpha_{t-1})x_{t-1}, \sigma_{t|t-1}^2 \mathbf{I}), \quad (5.1)$$

where $\mathcal{N}$ is a Gaussian distribution, $\mathbf{I}$ is the identity covariance matrix, and α , σ^2 are scaling and noise variance parameters where $\alpha_t^2 + \sigma_t^2 = 1$ [259]. In the reverse direction, a network parametrization $\mathbf{D}_\theta$ is used to restore original data from noisy samples, i.e., $\hat{x}_0 = \mathbf{D}_\theta(x_t, t)$. The prior can be trained by minimizing a variational bound on likelihood [260]:

$$\mathbb{E}_{t \sim U[1, T], x_0 \sim p(x), x_t \sim q(x_t|x_0)} [\omega(\lambda_t) \|\mathbf{D}_\theta(x_t, t) - x_0\|_2^2], \quad (5.2)$$

where $\mathbb{E}$ is expectation, U is a uniform distribution, $q(x_t|x_0) = \mathcal{N}(x_t; \alpha_t x_0, \sigma_t^2 \mathbf{I})$, $\omega(\lambda_t)$ is a weighting function with $\lambda_t = \log(\alpha_t^2/\sigma_t^2)$ denoting signal-to-noise ratio.

Once trained, the diffusion prior can generate a synthetic data sample by progressively denoising a random noise sample across T steps. The reverse transition probability for the denoising steps can be expressed as [259]:

$$q(x_{t-1}|x_t, \hat{x}_0) = \mathcal{N}\left(\alpha_{t-1}\hat{x}_0 + \sqrt{1 - \alpha_{t-1}^2 - \beta_t^2} \cdot \frac{x_t - \alpha_t \hat{x}_0}{\sigma_t}, \beta_t^2 \mathbf{I}\right), \quad (5.3)$$

where $\beta_t = \frac{\sigma_{t-1}}{\sigma_t \alpha_{t-1}} \sqrt{(\alpha_{t-1}^2 - \alpha_t^2)}$ controls the stochasticity of generated samples. The original diffusion prior requires $T \approx 1000$ forward passes through the network for generation. To lower sampling time, a common solution is interleaved

sampling with step size k , $x_{t-k} \sim q(x_{t-k}|x_t, \hat{x}_0)$, while $\beta_t = 0$ for deterministic generation (i.e., DDIM) [265]. Yet, interleaved sampling can still require few hundred steps for generation and it typically suffers from reduced sample quality [281].

5.3.2 DreaMR

DreaMR is a novel explanation method for deep fMRI models based on counterfactual generation. Assume that a downstream classifier $c(x)=y$ maps an input fMRI sample $x_0 \sim p(x)$ onto a class label $y_0 \in Y$ for cognition-related variable, according to posterior probability $p_c(y|x)$. Here, we take that $x_0 \in \mathbb{R}^{R \times W}$ denotes the BOLD responses recorded in the given fMRI scan, where R is the number of brain regions and W is the number of time frames. Counterfactual generation aims to obtain plausible samples $\bar{x}_0 \sim p(x)$ with minimal alterations from x_0 , such that the classifier decision is flipped $c(\bar{x}_0) = \bar{y}_0$ where $\bar{y}_0 \neq y_0, \bar{y}_0 \in Y$. Afterwards, the differences between the original and counterfactual samples ($x_0 - \bar{x}_0$) can be inspected to infer the input features that are critical in distinguishing between labels y_0 and $\bar{y}_0$.

Counterfactual generation inherently requires alteration of response values within the input fMRI sample. To do this based on a trained diffusion prior, a noisy fMRI sample $\bar{x}_{\Delta T}$ is first obtained by adding a moderate level of white Gaussian noise onto the original sample x_0 [280]:

$$q(\bar{x}_{\Delta T}|x_0) = \mathcal{N}(\bar{x}_{\Delta T}; (\alpha_{\Delta T}/\alpha_0)x_0, \sigma_{\Delta T|0}^2 \mathbf{I}), \quad (5.4)$$

where the time step $\Delta T < T$ is a hyperparameter. Starting reverse diffusion at ΔT , response values in the noisy fMRI sample can then be altered according to the reverse transition probabilities in Eq. 5.3, so that the resultant counterfactual sample remains proximal to x_0 . Yet, to ensure that the counterfactual sample is able to flip the classifier decision, classifier guidance is also injected [281]:

$$q(\bar{x}_{t-1}|\bar{x}_t, \hat{x}_0) = \mathcal{N}(\hat{x}_0 + s\beta_t^2 \nabla_{\bar{x}_t} \log p_c(\bar{y}_0|\bar{x}_t), \beta_t^2), \quad (5.5)$$

where $\nabla_{\bar{x}_t}$ is gradient with respect to $\bar{x}_t$, s is a scaling parameter that controls the relative weight of classifier guidance.

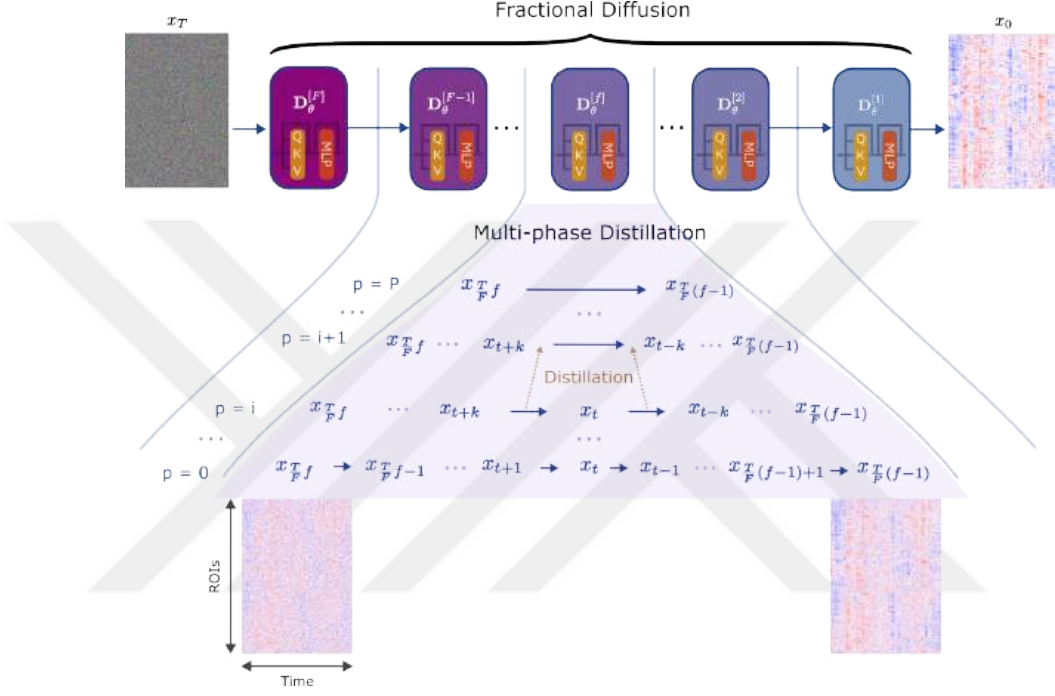


Figure 5.2: DreaMR leverages a novel FMD prior for efficient sample generation without compromising sample quality, and to enable effective intermittent control over the generation process. The FMD prior splits the overall diffusion process into F fractions, where the f th fraction covers T/F consecutive steps from $t_s(f) = T(f)/F$ to $t_e(f) = T(f-1)/F + 1$. Since the characteristics of the denoising task can vary notably across fractions, a dedicated denoising network $D_{\theta_P}^{[f]}$ is employed in each fraction to improve sample fidelity. To enhance sensitivity to long-range temporal context in fMRI scans, the network is built on an efficient transformer architecture that uses fused window attention mechanisms [1]. To improve sampling efficiency, multi-phase distillation is performed over P phases in each fraction (Eq. 5.7). In the p th distillation phase, the diffusion step size is doubled to shorten sampling time by a factor of 2. At the end of multi-phase distillation, the number of sampling steps that must be executed for a given fraction is reduced from T/F to $T/(2^P F)$. Note that the novel combination of diffusion fractions and multi-phase distillation in FMD serves to mitigate losses in sample quality typically encountered with distillation of conventional diffusion priors.

Previous studies on counterfactual generation have adopted interleaved sampling with conventional diffusion priors, which still takes few hundred sampling steps [265]. Furthermore, counterfactual generation requires knowledge of $\nabla_{\bar{x}_t} \log p_c(\bar{y}|\bar{x}_t)$ that is unknown a priori as the classifier is trained on original samples without added noise. To improve inference efficiency, DreaMR leverages a novel FMD prior for counterfactual generation in few steps without compromising sample quality (Fig. 5.1). Meanwhile, to avoid the need for classifier retraining on noisy samples, DreaMR computes classifier gradients given denoised sample estimates as a surrogate for gradients on noisy samples. Working principles of the FMD prior, classifier guidance and the counterfactual generation algorithm are detailed in the rest of this section.

5.3.2.1 Fractional multi-phase-distilled diffusion prior

During counterfactual generation with diffusion priors, providing guidance via classifier gradients at each diffusion step can cause significant computational burden for inference [261]. A mainstream approach to improve inference efficiency is distillation for post-hoc reduction of the number of diffusion steps [280, 283]. However, conventional distillation procedures on common diffusion priors can result in undesirable losses in sample quality [284, 285]. Here, we argue that two main contributors to these losses are the use of a single denoising network for the entire diffusion process and the use of a single-phase distillation. Thus, to enable efficient inference without compromising sample quality, we propose a novel FMD prior that splits the diffusion process into F uniform fractions with dedicated denoising networks, and performs multi-phase distillation separately in each fraction (Fig. 5.2).

Characteristics of the denoising task can show notable variations within a diffusion process due to varying noise levels and feature details across diffusion steps [286]. In turn, poor adaptation to these characteristics can induce significant performance losses in distilled diffusion priors. Several recent studies have considered to employ multiple denoising networks on separate time fractions to improve adaptation in undistilled diffusion priors [286], but these studies did

not examine the influence of fractional diffusion on distilled priors. While other studies have considered progressive distillation over multiple stages to alleviate losses in distilled priors [284], they employed a common denoising network across the entire diffusion process that can compromise sample quality. Unlike these recent efforts, here we introduce FMD as the first diffusion prior that synergistically combines fractional diffusion with fraction-specific multi-phase distillation to simultaneously maintain high quality and efficiency in sample generation.

For fractional diffusion, FMD employs a dedicated denoising network $\hat{x}_0 = \mathbf{D}_\theta^{[f]}(x_t, t)$ in each fraction (Fig. 5.2). The f th fraction covers T/F consecutive steps from $t_s(f) = T(f)/F$ to $t_e(f) = T(f-1)/F + 1$. The resultant FMD prior is trained via the following objective:

$$\sum_{f=1}^F \mathbb{E}_{t \sim U[t_s(f), t_e(f)]} \left[\|\mathbf{D}_\theta^{[f]}(x_t, t) - x_0\|_2^2 \right], \quad (5.6)$$

where ω and expectation over x_0, x_t are omitted for brevity.

For multi-phase distillation, FMD performs gradual knowledge transfer from the original teacher network $\mathbf{D}_{\theta_0}^{[f]}(x_t, t)$ onto a student network $\mathbf{D}_{\theta_P}^{[f]}$ over P phases. In the p th phase, $\mathbf{D}_{\theta_{p-1}}^{[f]}(x_t, t)$ is the teacher, $\mathbf{D}_{\theta_p}^{[f]}(x_t, t)$ is the student, and the diffusion step size is increased by a factor of 2 as follows:

$$\mathbb{E}_{t \sim U(\{t_s(f):T/(k_d F):t_e(f)\})} \left[\|\mathbf{D}_{\theta_p}^{[f]}(x_t, t) - \tilde{x}_0\|_2^2 \right], \quad (5.7)$$

where $k_d = 2^p$ is the diffusion step size of the student. Adopting interleaved sampling, the reference sample $\tilde{x}_0$ is derived via sampling based on the teacher $\mathbf{D}_{\theta_{p-1}}^{[f]}(x_t, t)$:

$$\tilde{x}_0 = \frac{\tilde{x}_{t-2k_o} - (\sigma_{t-2k_o}/\sigma_t)x_t}{\alpha_{t-2k_o} - (\sigma_{t-2k_o}/\sigma_t)\alpha_t}, \quad (5.8)$$

where $k_o = 2^{(p-1)}$ is the diffusion step size of the teacher. At the end of the distillation procedure, the number of steps for a given fraction is reduced from T/F to $T/(2^P F)$. Although this multi-phase distillation involves additional computations over a single-phase distillation during the training stage, progressively lowering the number of diffusion steps helps mitigate losses in sample quality in the distilled diffusion prior.

5.3.2.2 Classifier guidance

A counterfactual sample that flips the classifier decision can be drawn if the joint score function $\nabla_{x_t} \log p(x_t, y)$ for noisy samples and predicted class labels is known. Since $p(x_t, y) = p(x_t)p_c(y|x_t)$ based on Bayes' rule, the joint score is given as:

$$\nabla_{x_t} \log p(x_t, y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p_c(y|x_t). \quad (5.9)$$

The first term denotes the marginal score function for noisy samples, which can be derived as follows [259]:

$$p(x_t) = \frac{1}{\sigma_t \sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{x_t - \alpha_t \hat{x}_0}{\sigma_t} \right)^2 \right), \quad (5.10)$$

$$\nabla_{x_t} \log p(x_t) = -\frac{x_t - \alpha_t \hat{x}_0}{\sigma_t^2}. \quad (5.11)$$

Meanwhile, the second term denotes the conditional score function for predicted labels given noisy samples. Note that the originally trained classifier does not capture $p_c(y|x_t)$, but instead $p_c(y|x_0)$. The classifier could be retrained on a set of noisy samples generated by the diffusion prior. However, this brings in additional computational burden, and learning of $p_c(y|x_t)$ on samples with heavy noise might be difficult.

To avoid limitations related to classifier retraining, here we derive a surrogate based on $p_c(y|x_0)$ to compute the conditional score function. Following the chain rule:

$$\nabla_{x_t} \log p_c(y|x_t) = \nabla_{\tilde{x}_0} \log p_c(y|x_t) \cdot \nabla_{x_t} \tilde{x}_0, \quad (5.12)$$

where $\tilde{x}_0$ is obtained via sampling across the diffusion process. Based on forward diffusion, x_t and x_0 are related as:

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \text{ s.t. } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (5.13)$$

Assuming that $\tilde{x}_0 \simeq x_0$, $\nabla_{x_t} \tilde{x}_0 = 1/\alpha_t$. In conventional diffusion priors with stochastic sampling (i.e., $\beta_t \neq 0$), x_t and $\tilde{x}_0$ can be related through a one-to-many mapping. Yet, since DreamMR adopts deterministic sampling with $\beta_t = 0$, a

unique $\tilde{x}_0$ is obtained given x_t . As such, $p_c(y|x_t) = p_c(y|\tilde{x}_0)$, and the conditional score can be expressed as:

$$\nabla_{x_t} \log p_c(y|x_t) = \frac{1}{\alpha_t} \nabla_{\tilde{x}_0} \log p_c(y|\tilde{x}_0). \quad (5.14)$$

Since $p_c(y|\tilde{x}_0) \simeq p_c(y|x_0)$ for a reasonably well-trained diffusion prior, the posterior distribution of the originally trained classifier can be used to provide guidance without retraining.

5.3.2.3 Counterfactual generation

The counterfactual generation algorithm for DreaMR is outlined in Alg. 2. Given an original fMRI sample x_0 , a noisy sample $\bar{x}_{\Delta T}$ is first obtained by adding Gaussian noise via forward diffusion as described in Eq. 5.13. DreaMR then performs reverse diffusion sampling via the trained FMD prior to alter the response values in $\bar{x}_{\Delta T}$, without diverging significantly from x_0 . Note that time step ΔT corresponds to fraction $f_c = \lceil \Delta T(F/T) \rceil$, so reverse diffusion is initiated at the f_c th fraction. To ensure that the eventual counterfactual sample is able to flip the classifier decision to $\bar{y}$, gradient of the posterior distribution of the classifier is also employed. For the f th fraction, the interleaved sampling equation based on the FMD prior injected with classifier guidance can be described as:

$$\begin{aligned} \bar{x}_{t-k} = & \alpha_{t-k} \hat{\tilde{x}}_0 + \sigma_{t-k} \frac{\bar{x}_t - \alpha_t \hat{\tilde{x}}_0}{\sigma_t} + \\ & s \frac{\sigma_t^2}{\alpha_t^2} (\alpha_{t-k} - \alpha_t \frac{\sigma_{t-k}}{\sigma_t}) \nabla_{\tilde{x}_0} \log p_c(\bar{y}|\tilde{x}_0), \end{aligned} \quad (5.15)$$

where k is step size, $\hat{\tilde{x}}_0 = \mathbf{D}_{\theta_P}^{[f]}(\bar{x}_t, t)$, and s is a scaling constant to control the strength of guidance. In Eq. 5.15, the conditional score from the classifier is computed using $\tilde{\tilde{x}}_0$, which is Langevin sampled across the diffusion process, instead of $\hat{\tilde{x}}_0$, as this was observed to improve the quality of guidance. The counterfactual sample $\bar{x}_0$ is obtained after taking $\Delta T/k$ reverse diffusion steps.

Algorithm 2: Counterfactual generation with DreaMR

Input:

$x_0 \sim p(x)$: Original fMRI sample
 $p_c(y|x)$: Posterior probability of the classifier
 $y_0 \in Y$: Classifier-predicted label for x_0
 $\bar{y}_0$: Target label for counterfactual generation
 ΔT : Initial diffusion step for counterfactual generation
 $\{\mathbf{D}_{\theta_P}^{[1]}, \dots, \mathbf{D}_{\theta_P}^{[F]}\}$: Distilled networks across fractions
 $k = 2^P$: Step size after P distillation phases

Output:

$\bar{x}_0$: Counterfactual sample
 $\bar{x}_{\Delta T} \leftarrow \alpha_{\Delta T} x_0 + \sigma_{\Delta T} \epsilon$; $\triangleright$ generate noisy sample
for f_o *in* $\text{range}(f_c, 0, -1)$ **do**
 $\triangleright$ sample across fractions to find $\tilde{x}_0$
 for t_i *in* $\text{range}(t_s(f_o), -1, -k)$ **do**
 $f_i \leftarrow \text{ceil}(t_i F / T)$;
 $\hat{x}_0 \leftarrow \mathbf{D}_{\theta_P}^{[f_i]}(\bar{x}_{t_i}, t)$;
 $\bar{x}_{t_i-k} \leftarrow \alpha_{t_i-k} \hat{x}_0 + \sigma_{t_i-k} \frac{\bar{x}_{t_i} - \alpha_{t_i} \hat{x}_0}{\sigma_{t_i}}$;
 end
 $\tilde{x}_0 \leftarrow \bar{x}_0$; $G \leftarrow \nabla_{\tilde{x}_0} \log p_c(\bar{y} | \tilde{x}_0)$; $\triangleright$ gradient
 $\triangleright$ sample within fraction to find $\bar{x}_{t_e(f)}$
 for t_o *in* $\text{range}(t_s(f_o), t_e(f_o) - 1, -k)$ **do**
 $\hat{x}_0 \leftarrow \mathbf{D}_{\theta_P}^{[f_o]}(\bar{x}_{t_o}, t)$;
 $\gamma \leftarrow s \frac{\sigma_{t_o}^2}{\alpha_{t_o}^2} (\alpha_{t_o-k} - \alpha_{t_o} \frac{\sigma_{t_o-k}}{\sigma_{t_o}})$; $\triangleright$ scale
 $\bar{x}_{t_o-k} \leftarrow \alpha_{t_o-k} \hat{x}_0 + \sigma_{t_o-k} \frac{\bar{x}_{t_o} - \alpha_{t_o} \hat{x}_0}{\sigma_{t_o}} + \gamma G$;
 end
end

5.4 Methods

5.4.1 Experimental Procedures

Demonstrations were performed on fMRI scans from HCP-Rest, HCP-Task [103] and ID1000 datasets [210]. After exclusion of incomplete scans (<1200 s), HCP-Rest contained 1093 resting-state fMRI samples (594 female, 499 male). HCP-Task contained 7450 task-based fMRI samples (594 female, 501 male) where each subject performed 7 different tasks (i.e., emotion, relational, gambling, language,

social, motor, working memory). ID1000 contained 881 movie-watching fMRI samples (458 female, 423 male).

As downloaded from the public datasets, several preprocessing steps had already been performed on fMRI scans. For HCP-Rest and HCP-Task, motion correction, distortion correction, registration onto the MNI template, bias field correction, and brain extraction had been performed [287]. For ID1000, motion correction, distortion correction, registration onto the ICBM template, brain extraction, and CompCorr denoising had been performed [210]. In addition to these steps, we created nuisance variables for motion and physiological noise due to respiratory and cardiac traces across each fMRI run [45]. To control for confounds, these nuisance variables were regressed out from voxel-wise fMRI time series, and the resultant time series were z-scored across time for normalization.

Regions of interest (ROI) in the brain were defined based on the Schaefer atlas ($R=400$ regions) [150], the MMP atlas ($R=360$ regions) [288], or an aggregate atlas that pooled MMP with 20 cerebellar and 24 subcortical ROIs from the Talairach atlas ($R=404$) [289]. Responses within each ROI were obtained by averaging fMRI signals across member voxels according to the ROI definition. As such, each fMRI scan was represented via a data matrix of $R \times W$, with the number of time frames $W=1200$ for HCP-Rest, $W=176-405$ for HCP-Task depending on cognitive task, and $W=290$ for ID1000. To cope with varying scan durations, a sliding window approach was adopted across the time dimension of fMRI scans [1]. Counterfactual methods were devised to process individual strided windows across the scans, and window-specific outputs were merged by averaging overlapping time frames between consecutive windows. The window size was 600 for HCP-Rest, 128 for HCP-Task and ID1000, and the stride was taken as half the window size [1].

Experiments were conducted on single NVIDIA RTX 3090 GPUs using the PyTorch framework. Modeling was performed via a five-fold cross-validation procedure. In each fold, data were randomly split into training (80%), validation (10%) and test sets (10%) without any subject overlap among the three sets. Data splits were devised to ensure that there was no subject overlap among the

validation sets, or among the test sets for separate folds. For fair comparison, identical data splits were used for all competing methods.

A transformer-based downstream fMRI classifier was trained using cross-entropy loss [1]. For explanation methods, priors were trained using their originally proposed losses, and hyperparameters including number of epochs, learning rate and batch size were selected to minimize validation loss [126]. For each method, a common set of hyperparameters yielding near-optimal performance was used across datasets.

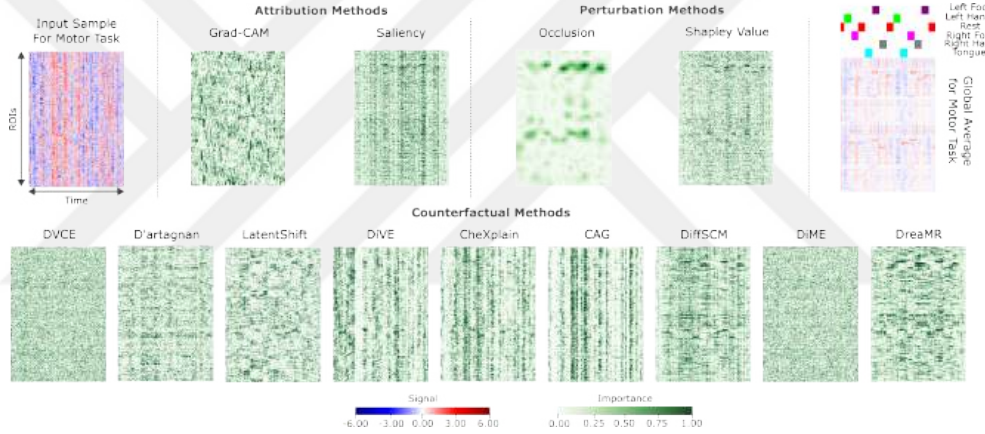


Figure 5.3: Spatiotemporal explanation maps from competing methods for a representative fMRI scan in HCP-Task based on the Schaefer atlas. The original input fMRI sample for the motor task (left) and the global average of fMRI samples across subjects for the motor task (right) are also shown. For attribution methods, explanation maps were taken as the gradient of the classifier loss function with respect to input features. For perturbation methods, explanation maps were obtained by masking out local patches in fMRI samples. For counterfactual methods, counterfactuals were generated separately to flip the class label from the motor onto each of six remaining cognitive tasks, and explanation maps were taken as the average difference between the original and counterfactual samples. See colorbars for fMRI samples (Signal) and explanation maps (Importance).

5.4.2 Competing Methods

DreaMR was compared against state-of-the-art counterfactual methods based on VAE, GAN, and diffusion priors.

5.4.2.1 DreaMR

DreaMR was implemented based on the transformer architecture in [1], adapted for diffusion modeling by incorporating time encoding via adaptive normalization layers [177]. Inspired by efficient transformer methods based on temporal windowing [290, 140], the transformer processed time series across a hierarchically growing set of time windows to maintain linear complexity. The FMD prior used $T=1024$ steps, $F=4$ fractions, $P=7$ distillation phases, $k=128$ final step size. Cross-validated hyperparameters were $E=100$ epochs, $\eta=2 \times 10^{-4}$ learning rate, $B=8$ batch size, $s=(8,16,32)$ for HCP-Task, ID1000, HCP-Rest respectively.

5.4.2.2 DiME

A diffusion-based method was trained to generate fMRI samples; counterfactuals were generated by interleaved sampling guided by downstream classification loss and perceptual loss between original and resampled fMRI scans [282]. $E=200$, $\eta=5 \times 10^{-5}$, $B=8$ were cross-validated.

5.4.2.3 DiffSCM

A diffusion-based method was trained to generate fMRI samples; counterfactuals were generated by interleaved sampling guided by downstream classification loss [281]. $E=200$, $\eta=10^{-4}$, $B=8$ were cross-validated.

5.4.2.4 CAG

A GAN-based method was trained to translate between classes given cross-entropy loss from the downstream classifier; counterfactuals were generated via translation [253]. $E=200$, $\eta=5 \times 10^{-5}$, $B=32$ were cross-validated.

5.4.2.5 CheXplain

A GAN-based method was trained to generate samples given predictions by the downstream classifier; counterfactuals were generated by modifying the style latents [291]. $E=200$, $\eta=2 \times 10^{-4}$, $B=16$ were cross-validated.

5.4.2.6 DiVE

A VAE-based method was trained via a variational objective to generate fMRI samples; counterfactuals were generated by modifying encoded latents of the autoencoder [254]. $E=500$, $\eta=4 \times 10^{-4}$, $B=16$ were cross-validated.

5.4.2.7 LatentShift

An autoencoder-based method was considered that modified latent representations of an input fMRI sample to emphasize features that contribute to the classifier decision [258]. $E=100$, $\eta=4 \times 10^{-4}$, $B=16$ were cross-validated.

5.4.2.8 D’artagnan

A GAN-based method was trained to translate between classes given cross-entropy loss from the downstream classifier and distance between the original and translated fMRI scans; counterfactuals were generated via translation [278]. $E=100$, $\eta=1 \times 10^{-4}$, $B=8$ were cross-validated.

5.4.2.9 DVCE

A diffusion-based method was trained to generate fMRI samples; counterfactuals were obtained by interleaved sampling guided by downstream classification loss [280]. $E=100$, $\eta=5 \times 10^{-5}$, $B=8$ were cross-validated.

Table 5.5: Fidelity of spatiotemporal and FC features of counterfactual samples for the HCP-Task dataset. Results based on an aggregate atlas combining ROIs from the MMP and Talairach atlases.

	Spatiotemporal			FC		
	Prox. ↓	Spar. ↓	FID ↓	Prox. ↓	Spar. ↓	FID ↓
DreaMR	56.5±0.4	17.6±0.2	4.9±0.1	1.0±0.0	5.1±0.2	6.1±0.4
DiME	69.4±0.7	21.7±0.2	9.6±1.3	1.5±0.1	10.5±1.1	14.1±1.1
DiffSCM	76.0±9.0	14.6±0.9	21.5±1.5	5.1±0.5	27.0±1.7	55.6±7.5
CAG	155.0±41.9	39.6±5.4	13.5±0.9	2.7±0.2	20.8±1.2	22.5±3.6
CheXplain	202.6±8.6	47.2±8.6	98.2±4.0	4.2±0.1	31.1±0.6	92.8±10.2
DiVE	107.1±0.9	32.7±0.2	129.1±5.9	23.6±1.2	88.2±1.3	182.4±6.0
LatentShift	122.2±55.7	31.3±9.3	123.6±14.5	11.5±2.2	58.5±7.8	141.4±15.1
D’artagnan	64.3±6.9	19.9±2.6	29.6±6.0	5.4±0.3	41.7±1.6	81.5±13.2
DVCE	57.9±0.8	17.8±0.3	19.9±2.1	2.4±0.2	18.6±2.4	26.1±3.8

5.4.3 Performance Evaluation

We quantified proximity, sparsity and Fréchet Inception Distance (FID) metrics to assess performance of competing counterfactual methods [280]. Proximity was taken as the normalized ℓ_2 distance between original and counterfactual samples; sparsity was taken as the proportion of features whose absolute difference between original and counterfactual samples exceeded the standard deviation across features in the original sample. Low proximity and sparsity indicate that feature changes between samples are minimal, so the resultant interpretations are specific. Meanwhile, low FID scores suggest that counterfactual samples are drawn from a similar distribution to that of original samples, so the interpretations are plausible. Significance of differences between methods were assessed across test subjects via non-parametric Wilcoxon signed-rank tests. FID produced an aggregate measure across the test set so it was not included in significance assessments.

5.5 Results

5.5.1 Fidelity of Counterfactual fMRI Samples

We first demonstrated DreaMR in explanation of downstream transformer-based classifiers for sex on HCP-Rest and ID1000, and for cognitive task on HCP-Task. Separate analyses were conducted using the Schaefer and MMP atlases to rule out potential concerns regarding bias in ROI definitions. Fig. 5.3 depicts representative explanation maps for spatiotemporal features of an fMRI sample from the motor task, produced by competing counterfactual as well as attribution [274] and perturbation [292, 293] methods. Attribution methods tend to yield broad-spread maps with poor specificity for local features in the fMRI sample. Perturbation methods can improve capture for some local features, but their results are inconsistent across the fMRI sample given limited sensitivity for global context. Meanwhile, counterfactual methods based on VAE, GAN priors tend to emphasize non-salient features for the motor task, suggesting suboptimal fidelity in counterfactual samples. In contrast to competing methods, DreaMR generates an explanation map that is more closely aligned with prominent spatiotemporal features in the average sample for the motor task, without signs of over-broadening or incoherence. These qualitative assessments suggest that DreaMR can offer high fidelity in counterfactual explanation.

To provide specific explanations, counterfactual generation aims to identify minimal changes in the features of an original fMRI sample that are sufficient to flip the respective classifier decision. Thus, under the condition of successful decision flipping, it is desirable that counterfactual and original samples follow similar data distributions. Based on this notion, the fidelity of counterfactual samples was quantitatively evaluated via proximity, sparsity and FID metrics (see Section 5.4.3 for definitions), which reflect the similarity between the counterfactual and original data distributions. Proximity, sparsity and FID values for spatiotemporal features of counterfactual samples are listed in Table ?? based on the Schaefer atlas, and in Table ?? based on the MMP atlas. Note that, in these quantitative evaluations, each competing method yielded successful counterfactuals that were

able to flip the classifier decisions for all original samples in the test sets. In general, the relative performance levels of competing methods show similar trends for analyses based on Schaefer versus MMP atlases, suggesting that our results are not unduly biased by ROI definitions. Overall, DreaMR achieves the lowest proximity, sparsity and FID across datasets and atlases ($p < 0.05$), except for DiffSCM that yields moderately lower sparsity on HCP-Task and ID1000 datasets, and CAG that yields moderately lower FID on HCP-Rest and on ID1000 with the MMP atlas. Note that, in these exception cases, DreaMR still performs competitively with DiffSCM or CAG, and yields the second-best performance among competing methods. On average across atlases, DreaMR outperforms competing methods in proximity by 51.6, sparsity by 13.9, FID by 57.6 on HCP-Rest; proximity by 49.7, sparsity by 11.9, FID by 47.8 on HCP-Task; and proximity by 70.4, sparsity by 16.7, FID by 57.8 on ID1000. These results indicate that DreaMR generates high-fidelity counterfactuals minimally different from the original fMRI samples and closely aligned with the original data distribution in terms of spatiotemporal features.

In the neuroimaging literature, another pervasively analyzed attribute of fMRI data are functional connectivity (FC) features that are commonly associated with cognition-related variables [82]. Thus, we also examined the fidelity of FC features derived from counterfactual fMRI samples. To do this, a counterfactual sample was generated via competing methods on each original fMRI sample. For both counterfactual and original samples, the FC feature between a given pair of brain regions was taken as Pearson’s correlation coefficient between respective fMRI time courses [90]. Note that if a particular method yields counterfactual samples that are more similar to the original fMRI samples, then the FC features derived from those counterfactuals are likely to be more similar to the FC features of original samples. Thus, in terms of relative performance among competing methods, results for FC features are expected to follow partly similar patterns with those for spatiotemporal features. Proximity, sparsity and FID values for FC features are listed in Table ?? based on the Schaefer atlas, and in Table ?? based on the MMP atlas. The relative performance levels of competing methods are generally similar based on Schaefer versus MMP atlases, corroborating that our results are

not unduly biased by ROI definitions. Overall, DreaMR achieves the top performance in all datasets and atlases ($p < 0.05$). On average across atlases, DreaMR outperforms competing methods in proximity by 8.6, sparsity by 42.9, FID by 99.7 on HCP-Rest; proximity by 6.5, sparsity by 35.5, FID by 68.4 on HCP-Task; and proximity by 5.4, sparsity by 29.6, FID by 103.6 on ID1000. These results indicate that DreaMR generates reliable counterfactual samples whose FC features are closely aligned to the distribution of FC features for original fMRI samples, similar to the results on spatiotemporal features. Note, however, that FC features reflect time-aggregated connectivity measures that typically follow a lower dimensional distribution compared to spatiotemporal features, and the distributional attributes of the two feature sets show notable differences [82]. As such, a simple comparison of fidelity metrics for FC versus spatiotemporal features might yield misleading impressions regarding relative success in counterfactual explanation, and further work is warranted to elucidate this challenging question.

Certain cognitive tasks might involve brain regions not only within but also outside the cerebral cortex, for which the Schaefer and MMP atlases provide a tessellation [294]. For instance, it has been suggested that motor and working memory tasks evoke responses in cerebellar regions, whereas gambling tasks can evoke responses in subcortical regions [103]. To provide a broader assessment across the brain, we conducted a separate analysis based on an aggregate atlas that combined ROI definitions for cortical regions from the MMP atlas with those for cerebellar and subcortical regions from the Talairach atlas. A downstream transformer-based classifier was first trained to detect cognitive tasks on the HCP-Task dataset, and competing methods were then employed for counterfactual explanation. Proximity, sparsity and FID values for both spatiotemporal and FC features of counterfactual samples are listed in Table 5.5. For both types of features, we find that the fidelity metrics of competing methods show very similar distributions based on MMP versus aggregate atlases, where DreaMR generally yields the lowest metric values among competing methods. These results suggest that DreaMR generates high-quality counterfactuals that are closely aligned with the distribution of original fMRI samples not only in cortical but also in cerebellar and subcortical regions.

Table 5.6: Detection performance of a linear classifier trained using counterfactual samples was measured via accuracy (Acc.) and F1 metrics. Results listed for HCP-Rest, HCP-Task and ID1000 based on the Schaefer atlas.

	HCP-Rest		HCP-Task		ID1000	
	Acc. $\uparrow$	F1 $\uparrow$	Acc. $\uparrow$	F1 $\uparrow$	Acc. $\uparrow$	F1 $\uparrow$
DreaMR	78.0$\pm$2.1	80.0$\pm$2.3	92.7$\pm$3.0	92.8$\pm$3.0	78.0$\pm$2.4	75.8$\pm$5.2
DiME	47.8 $\pm$ 2.7	63.2 $\pm$ 1.6	84.6 $\pm$ 2.5	83.8 $\pm$ 2.9	68.9 $\pm$ 6.4	74.9 $\pm$ 4.3
DiffSCM	53.0 $\pm$ 1.7	63.0 $\pm$ 1.2	88.8 $\pm$ 4.4	89.2 $\pm$ 3.9	52.6 $\pm$ 3.5	53.1 $\pm$ 1.1
CAG	62.8 $\pm$ 11.6	70.8 $\pm$ 5.1	69.0 $\pm$ 5.7	69.1 $\pm$ 6.4	57.50 $\pm$ 8.4	69.8 $\pm$ 5.0
CheXplain	52.6 $\pm$ 2.0	64.2 $\pm$ 0.4	19.6 $\pm$ 5.9	14.1 $\pm$ 4.7	53.0 $\pm$ 3.7	67.1 $\pm$ 1.8
DiVE	64.9 $\pm$ 10.6	71.1 $\pm$ 5.3	64.5 $\pm$ 10.9	63.1 $\pm$ 11.7	66.1 $\pm$ 9.1	74.2 $\pm$ 4.8
LatentShift	52.1 $\pm$ 4.3	65.6 $\pm$ 3.1	76.9 $\pm$ 3.1	76.3 $\pm$ 3.9	51.5 $\pm$ 3.2	66.6 $\pm$ 1.8
D’artagnan	55.3 $\pm$ 6.6	47.4 $\pm$ 17.4	56.5 $\pm$ 6.3	53.9 $\pm$ 6.4	49.1 $\pm$ 2.3	40.0 $\pm$ 29.9
DVCE	58.5 $\pm$ 7.6	68.5 $\pm$ 3.9	76.1 $\pm$ 12.8	74.7 $\pm$ 14.2	60.5 $\pm$ 9.0	71.1 $\pm$ 6.8

Table 5.7: Detection performance of a linear classifier trained using counterfactual samples for HCP-Task based on the aggregate atlas. Results listed separately for cortical and non-cortical ROIs.

	Cortical		Non-cortical	
	Acc. $\uparrow$	F1 $\uparrow$	Acc. $\uparrow$	F1 $\uparrow$
DreaMR	90.6$\pm$2.0	90.4$\pm$2.0	24.9$\pm$2.5	21.3$\pm$3.8
DiME	83.3 $\pm$ 5.7	83.1 $\pm$ 5.9	17.3 $\pm$ 1.8	12.3 $\pm$ 3.2
DiffSCM	68.7 $\pm$ 6.8	66.0 $\pm$ 6.0	19.8 $\pm$ 1.9	12.8 $\pm$ 1.9
CAG	66.5 $\pm$ 6.7	65.6 $\pm$ 7.0	21.5 $\pm$ 3.8	15.5 $\pm$ 6.0
CheXplain	21.0 $\pm$ 5.3	13.7 $\pm$ 4.3	14.1 $\pm$ 0.5	6.4 $\pm$ 1.6
DiVE	82.9 $\pm$ 3.0	82.8 $\pm$ 3.2	20.9 $\pm$ 3.4	12.2 $\pm$ 4.3
LatentShift	82.1 $\pm$ 2.5	81.9 $\pm$ 2.7	19.7 $\pm$ 2.0	16.7 $\pm$ 2.8
D’artagnan	62.9 $\pm$ 8.1	61.8 $\pm$ 7.0	16.9 $\pm$ 0.7	10.6 $\pm$ 3.3
DVCE	19.1 $\pm$ 4.3	11.4 $\pm$ 5.2	15.0 $\pm$ 0.4	5.0 $\pm$ 0.8

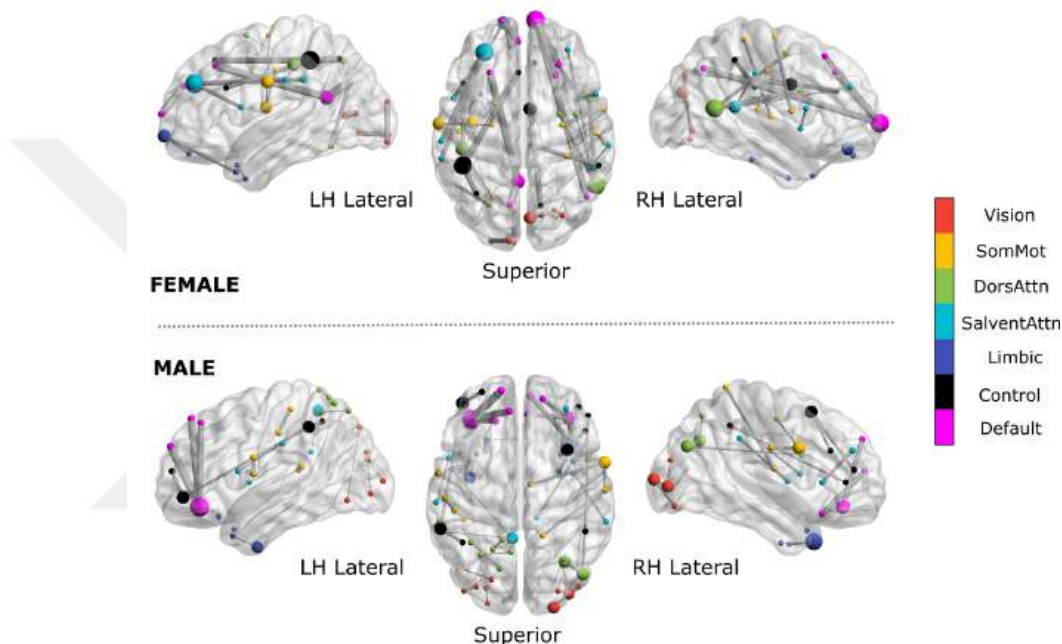


Figure 5.4: Cortical explanation maps produced by DreaMR for female (top panel) and male (bottom panel) sexes on the HCP-Rest dataset. To obtain the cortical maps, counterfactual samples were generated for each original fMRI sample to flip the respective decision of a deep fMRI classifier for sex detection. FC features of original and counterfactual samples were separately derived, and the differences between the two sets of samples were averaged across subjects. Important FC features were determined by selecting the features showing the top 5% of differences. Each important region-of-interest (ROI) in the brain is marked with a dot on the anatomical template, and connectivity between ROIs is shown with a bar. Dot size denotes the importance of the ROI, and bar thickness denotes the importance of the connection. ROIs are colored according to the functional network they belong to (see legend). Lateral and superior brain views are displayed. LH: left hemisphere, RH: right hemisphere; SomMot: somatomotor; DorsAttn: dorsal attention; SalventAttn: Salience/ventral attention.

5.5.2 Explanations based on Counterfactual Samples

Neuroscience studies routinely report that individual cognition-related variables (e.g., sex or cognitive task) are associated with characteristic patterns of brain responses [82, 45]. Given a downstream deep-learning classifier trained to detect such variables, the goal of counterfactual explanation is to render important input features explicitly observable in the differences between counterfactual and original fMRI samples, such that classifier decisions can be directly interpreted without the need for further processing. This bears out that the differences between counterfactual and original fMRI samples should ideally follow the differences between characteristic response patterns for the target versus original variables. As such, the explanatory capability of a counterfactual method can be assessed by examining the discriminative information that the counterfactual samples explicitly carry regarding the target versus original variables. Note that nonlinear classifiers (e.g., transformer-based classifiers explained here) subject their inputs to multiple levels of processing to unravel hidden features that are not directly observable in their inputs, and so their results can reflect spurious bias from the architecture of the classification model rather than the explicit information content of the inputs [295, 82]. To avoid potential biases, we adopted a linear classifier to assess the discriminative information that is explicitly observable in counterfactual samples [296]. A separate linear classifier was fit using the counterfactual samples generated by each competing method [1]. The linear classifiers were then tested on the original fMRI samples to detect the same variables as the nonlinear transformer-based classifiers. Table 5.6 lists accuracy and F1 for linear classifiers based on competing methods and the Schaefer atlas. DreaMR elicits the highest detection performance in all cases ($p < 0.05$). On average, DreaMR outperforms competing methods in accuracy by 22.1%, F1 by 15.9% on HCP-Rest; accuracy by 25.7%, F1 by 27.3% on HCP-Task; accuracy by 20.6%, F1 by 11.2% on ID1000. This finding indicates that DreaMR generates counterfactual samples that better capture the characteristic differences in response patterns between distinct cognition-related variables. We also questioned whether this benefit is evident beyond the cerebral cortex. For this purpose, we examined the discriminative

information in counterfactual samples for HCP-Task based on the aggregate atlas including non-cortical ROIs. Each counterfactual sample generated based on the aggregate atlas was split into two subsamples, one containing only cortical ROIs and the other containing only non-cortical ROIs. Accuracy and F1 metrics of separate linear classifiers fit for cortical versus non-cortical ROIs are listed in Table 5.7. Among competing methods, DreaMR attains the highest detection performance for both cortical and non-cortical ROIs ($p < 0.05$). Note that while non-cortical ROIs generally show above chance-level detection performance (i.e., $> 14.3\%$ accuracy), the detection performance for cortical ROIs is substantially higher. This finding implies that fMRI responses in cortical regions represent a greater amount of information about cognitive tasks included in the HCP-Task dataset than those in non-cortical regions.

Next, counterfactual samples from DreaMR were analyzed to derive cortical explanation maps that reflect the important FC features associated with individual classes detected by downstream transformer-based classifiers. To do this, differences of FC features between original and counterfactual samples were computed. Important features showing the top 5% of differences across subjects were determined. As a representative case, cortical explanation maps for each sex on HCP-Rest are displayed in Fig. 5.4 based on the Schaefer atlas. For females, relatively important features are observed in dorso-medial prefrontal and medial-posterior segments of the default mode network, parietal and medial segments of the control network, orbitofrontal segments of the limbic network, temporal-occipital and prefrontal segments of the salience/ventral attention network, posterior segments of the dorsal attention network, central segments of the somatomotor network, and medial segments of the vision network. Meanwhile, for males, important features are observed in ventral- and dorsal-prefrontal segments of the default mode network, lateral-parietal and lateral-prefrontal segments of the control network, temporal-pole segments of the limbic network, medial segments of the salience/ventral attention network, posterior segments of the dorsal attention network, precentral segments of the somatomotor network, and lateral segments of the vision network. Note that these patterns in the cortical distribution of important FC features are closely aligned with the neuroimaging

literature that identifies prominent sex-related differences in resting-state connectivity across the default mode, control, limbic, attention and somatomotor networks [156, 128]. In particular, ROIs distributed across relatively medial segments of default mode and control networks, associated with social cognition and executive control, have been suggested to show relatively higher FC values in females than males [297]. Similarly, ROIs in relatively orbitofrontal segments of the limbic network associated with emotion and memory have been reported to show higher FC values in females than males [298, 299]. While both sexes can show important FC features in the somatosensory network associated with motor tasks and in attention networks associated with multi-sensory spatial tasks, FC values across these networks have been reported to show stronger tendency for left lateralization in females and right lateralization in males [300, 297], as also evident in our cortical explanation maps. The overall consistency of our findings with the neuroimaging literature suggests that DreaMR is a promising framework to identify fMRI features that are characteristic to individual variables such as sex that influence cognitive function, and hence drive the decisions of downstream fMRI classifiers for these variables.

Table 5.8: Inference times (Inf., msec) and memory load (Mem., gigabytes) per generation of a counterfactual fMRI sample.

	HCP-Rest		HCP-Task		ID1000	
	Inf.	Mem.	Inf.	Mem.	Inf.	Mem.
DreaMR	2735	3.7	1173	2.7	1114	2.7
DiME	22104	6.0	14806	4.2	14918	4.2
DiffSCM	14327	3.5	7049	2.3	7343	2.4
CAG	116	4.3	45	3.2	45	3.3
CheXplain	384	3.0	491	2.9	282	2.9
DiVE	1750	3.8	1121	3.1	1149	2.8
LatentShift	967	3.7	850	2.9	829	2.2
D’artagnan	146	3.3	62	3.3	59	3.3
DVCE	23456	5.1	10740	4.1	10293	4.1

5.5.3 Computational Efficiency

A practical concern for counterfactual generation is the efficiency in resampling of original fMRI samples. Table 5.8 lists inference time and memory use of competing methods based on the Schaefer atlas. In inference time, non-iterative CAG and D’artagnan methods are the fastest, whereas iterative DiME, DiffSCM and DVCE based on conventional diffusion priors are notably slower. The remaining iterative methods including DreaMR require intermediate inference times between the two extremes. Note that the FMD prior in DreaMR enables substantially improved efficiency over conventional diffusion priors, and comparable efficiency with VAE and GAN priors. In terms of memory load, DreaMR has comparable demand to methods based on VAE and GAN priors, comparable demand to DiffSCM, and moderately lower demand than DiME.

Table 5.9: Fidelity of spatiotemporal features of counterfactual samples generated by DreaMR variants formed without a transformer, without Langevin sampling, without fractional diffusion, without multi-phase distillation, and with guidance from classifiers trained on noisy fMRI samples.

	Prox. ↓	Spar. ↓	FID ↓
DreaMR	41.4±0.5	11.7±0.2	21.2±0.8
w/o transformer	48.2±0.6	14.5±0.2	23.8±1.0
w/o Langevin	43.9±0.2	12.8±0.1	20.5±1.0
w/o fraction	42.5±1.0	12.2±0.4	21.1±1.1
w/o mp distillation	47.9±0.7	14.5±0.3	22.1±1.3
w noisy-classifier	41.6±0.6	11.8±0.2	21.3±1.1

Table 5.10: Fidelity of spatiotemporal features of counterfactual samples for DreaMR variants formed by implementing the FMD prior with varying number of fractions (F) and number of distillation phases (P).

Fractions				Distillation phases			
F	Prox.↓	Spar. ↓	FID ↓	P	Prox.↓	Spar. ↓	FID ↓
1	42.5±1.0	12.2±0.4	21.1±1.1	5	45.1±0.5	13.3±0.2	19.7±1.1
2	41.3±0.7	11.7±0.3	21.0±0.8	6	43.6±0.5	12.7±0.2	20.1±0.6
4	41.4±0.5	11.7±0.2	21.2±0.8	7	41.4±0.5	11.7±0.2	21.2±0.8
8	41.0±0.4	11.6±0.2	21.8±0.5	8	44.9±0.6	13.2±0.3	19.7±1.1

5.5.4 Ablation Studies

A series of ablation studies were conducted on the HCP-Rest dataset to assess the importance of the main design elements in DreaMR. For this purpose, several ablated variants of DreaMR were considered. First, we examined the influence of utilizing a transformer architecture, calculating the denoised sample estimate for classifier guidance based on iterated Langevin sampling, using fractional diffusion, using multi-phase distillation, and using classifier gradients on denoised

sample estimates. A ‘w/o transformer’ variant replaced the transformer with the common UNet architecture for diffusion priors [265]. A ‘w/o Langevin’ variant replaced the denoised sample estimate obtained through iterated Langevin sampling across diffusion fractions with a single-shot estimate predicted by the denoising network of the current fraction. A ‘w/o fraction’ variant used a single fraction for the diffusion prior. A ‘w/o mp distillation’ variant used a single-phase distilled diffusion prior. A ‘w noisy-classifier’ variant computed the conditional score function for classifier guidance by training a separate classifier on noisy fMRI samples instead of using the original classifier on clean samples. Proximity, sparsity and FID values for spatiotemporal features of counterfactual samples are listed in Table 5.9 based on the Schaefer atlas. Overall, we find that DreaMR outperforms all ablated variants, except for ‘w/o Langevin’ that yields moderately lower FID and ‘w noisy-classifier’ that generally performs similarly. Compared to DreaMR, ‘w/o transformer’ yields 6.8 higher proximity (16.4% performance loss), 2.8 higher sparsity (23.9% loss), 2.6 higher FID (12.3% loss); ‘w/o Langevin’ yields 2.5 higher proximity (6.0% loss), 1.1 higher sparsity (9.4% loss); ‘w/o fraction’ yields 1.1 higher proximity (2.7% loss), 0.5 higher sparsity (4.3% loss), and 0.1 higher FID (0.5% loss); ‘w/o mp distillation’ yields 6.5 higher proximity (15.7% loss), 2.8 higher sparsity (23.9% loss), 0.9 higher FID (4.2% loss). While diffusion-based elements elicit relatively limited benefits for FID, their contributions to proximity and sparsity are more comparable to the transformer architecture. Note that, in calculating the score function for classifier guidance, the ‘w noisy-classifier’ variant replaces the surrogate gradients based on the original classifier with the true gradients of a classifier separately trained on noisy fMRI samples. We observe that DreaMR and ‘w noisy-classifier’ perform very similarly, suggesting that the gradients of the original classifier serve as a successful surrogate. Taken together, these results suggest that each interrogated design element contributes significantly to method performance.

We then examined the influence of the number of diffusion fractions (F) and distillation phases (P) on the fidelity of the counterfactual samples generated by the FMD prior. To do this, variants of DreaMR were trained for varying F while $P=7$, and for varying P while $F=4$. Table 5.10 lists proximity, sparsity

and FID values for spatiotemporal features of counterfactual samples produced by DreaMR variants based on the Schaefer atlas. We find that the selected values of $F=4$, $P=7$ generally attain near-optimal performance. While $F=8$ yields slightly lower proximity and sparsity values, $F=4$ is preferable as it offers higher training efficiency by halving the number of distinct denoising networks to be learned. While $P=5$ and $P=8$ yield slightly lower FID, they have moderately higher proximity and sparsity values, and $P=7$ offers over sixteen times faster sampling than $P=5$ since inference time for counterfactual generation scales nearly quadratically with the number of diffusion steps.

5.6 Discussion

DreaMR is a counterfactual explanation method for fMRI based on a novel FMD prior that generates high-quality counterfactuals. Experimental demonstrations indicate the superior performance of DreaMR against competing counterfactual methods based on VAE, GAN and conventional diffusion priors. In particular, the difference between original and counterfactual samples produced by DreaMR enables more reliable explanations than baselines. At the same time, DreaMR achieves significantly improved efficiency in sample generation over conventional diffusion priors, resulting in brief inference times that approach the repetition time (TR) of common fMRI acquisitions.

The inference efficiency of DreaMR can be beneficial in real-time fMRI applications such as intraoperative planning, brain-computer interfaces, communication with locked-in patients, and therapeutic regulation of brain activations [266]. Real-time fMRI requires rapid detection of cognition-related variables to provide timely feedback for user intervention (e.g., a surgeon deciding on resection margins while the subject performs a cognitive task, or a patient trying to control brain activity to suppress tremor). Deep-learning classifiers promise sensitive detection of cognitive variables from real-time fMRI data, albeit poor interpretability can limit user trust in classifier decisions [301]. With its fast inference, DreaMR can facilitate rapid intervention by avoiding undesired delays in

counterfactual explanation. Another domain that can benefit from the efficiency of DreaMR is cohort studies that involve large-scale analyses of fMRI scans from a subject population to associate features in imaging data with specific classes of neurological disease [267]. Although deep-learning classifiers have emerged as state-of-the-art tools in establishing such associations, they still suffer from limited interpretability [267]. Note that explaining deep-learning classifiers with conventional diffusion priors can be computationally burdening for fMRI datasets comprising thousands of subjects, tens of different disease classes, and various different candidates for classifier architectures. In this context, adopting DreaMR for explaining classifier decisions can help improve reliability and efficiency of data analyses in cohort fMRI studies.

Limitations

Several lines of technical limitations could be addressed to further improve DreaMR. Following common practice, here we analyzed fMRI data preprocessed to register brain volumes onto an anatomical template. This procedure facilitates comprehensive and consistent region definitions across subjects based on an atlas [249], yet registration onto a template can yield spatial information losses. Potential losses might be mitigated by backprojecting atlas-based region definitions onto the brain spaces of individual subjects [302], and preserving spatial representations of individual subjects via enhanced localization mechanisms in transformer models [303, 233, 304].

The diffusion priors in DreaMR were trained here from scratch on public fMRI datasets comprising several hundred subjects. Literature suggests that such sizable datasets might be critical in adequate training of diffusion priors. In resource-limited application domains where data are scarce, fine tuning diffusion priors pretrained on time-series data such as audio waveforms might help improve learning [305]. Knowledge distillation from adversarial priors might also be employed to facilitate training of diffusion priors [?, 306, 307].

Counterfactual generation with a diffusion prior requires formation of an initial sample that reflects a noisy representation of the original data sample. In this study, initial samples were derived by adding random Gaussian noise onto the original samples, following the forward diffusion process. It may be possible to further improve similarity between counterfactual and original samples using alternative techniques such as DDIM inversion [281]. The inversion method uses estimates of noise components in the data samples to further corrupt them in order to obtain noisy representations. This can help increase similarity between counterfactual and original samples by improving preservation of non-noise data features [281]. However, it might also elevate noise correlations between counterfactual and original samples, potentially restricting the representational capacity of the diffusion prior. In unreported analyses, we observed that the inversion method can lower the amount of discriminative information about cognitive variables captured by counterfactual samples. Yet, further work is warranted to comprehensively evaluate the relative benefits of random noise addition versus DDIM inversion in the context of counterfactual explanations.

Here, DreaMR was implemented with uniform diffusion fractions of equal duration to give similar emphasis on each fraction. Alternatively, non-uniform fractions could be employed to help account for potential variability in the denoising task across the diffusion process due to varying noise levels and feature details. When a sufficiently large number of fractions are prescribed that help effectively capture task variability, we would not expect substantial differences between uniform versus non-uniform fractions. Yet, when a relatively limited number of fractions are prescribed, non-uniform fractions might offer potential performance benefits. In such cases, longer-duration fractions can be employed in diffusion segments of relatively slow variation, and shorter-duration fractions can be employed in segments of relatively fast variation in the characteristics of the denoising task. Future work is warranted to systematically assess the performance benefits of non-uniform versus uniform diffusion fractions.

To shorten sampling times for counterfactual generation, DreaMR leverages multi-phase distillation of its fractional diffusion prior. This distillation procedure lowers computational burden for the testing stage, at the expense of additional

burden introduced by distillation in the training stage. Corroborating recent reports, here we observed that distillation via a moderate number of phases helps improve sample quality compared to using a relatively limited number of phases [284]. Still, when needed, single-phase distillation procedures can be adopted to improve efficiency during the training stage at the expense of moderate losses in sample quality. Here, we obtained efficient explanations for fMRI samples lasting hundreds of time frames. Naturally, the burden of counterfactual generation during the testing phase grows with the temporal dimension of fMRI scans. In applications where the load becomes excessive due to high temporal resolution or long scan duration, the FMD prior might be combined with other efficient sampling approaches for accelerated diffusion [177, 308, 20]. During the testing stage, DreaMR uses a nested algorithm where an inner loop produces estimates for the clean sample via Langevin sampling, and this estimate is then used to compute classifier guidance that drives the outer loop to produce the counterfactual sample. This algorithm was adopted as the inner loop does not induce undue burden given the large step sizes prescribed for DreaMR, and since it was observed to yield higher sample quality. That said, the efficiency and simplicity of the sampling algorithm might be improved by discarding the inner loop and instead using the single-shot network estimate for the clean sample. Further work is warranted to systematically assess the benefits of various sampling algorithms in terms of sample fidelity versus efficiency during counterfactual generation.

Future Work

Here, we utilized DreaMR to explain transformer-based classifiers that predict discrete cognitive states given brain responses. Several extensions can be pursued to expand the scope of the proposed methodology. First, classifiers based on alternative convolutional or recurrent architectures might be considered [91, 126, 128]. Since counterfactual generation is a model-agnostic framework, DreaMR can in principle be adopted to other architectures without modification. Second, many neurodegenerative and neurodevelopmental diseases are reported to have complementary biomarkers in fMRI as well as anatomical or diffusion-weighted

MRI [309, 310]. A more performant classifier for disease-related cognitive states could be attained by using multi-modal images as input. DreaMR can be adopted for explaining such multi-modal classifiers by training a multi-modal FMD prior.

DreaMR also holds potential for explaining regression models in fMRI analysis based on deep neural networks. The human brain represents information on continuous stimulus or task variables, which can be decoded from brain responses via regression [45]. Deep regression models can also be employed to predict voxel- or ROI-wise responses given stimulus variables [168]. To adopt DreaMR, guidance from classification loss during counterfactual generation could be replaced with guidance from regression losses. It remains important future work to assess the efficacy of DreaMR in a broader set of explanation tasks in fMRI analysis.

It might be possible to adopt DreaMR to explain deep-learning models used to analyze other types of neuroimaging data beyond fMRI scans. Neuroimaging studies principally record spatiotemporal features of neural activity while cognitive variables are experimentally manipulated. For instance, electroencephalography measures local field potentials [311] and magnetoencephalography measures local magnetic fields caused by neural activity [312], near-infrared spectroscopy measures hemodynamic changes consequent to neural activity [313], and micro-electrode arrays directly measure multi-unit activity [314]. Inferences are then drawn by analyzing the association between measured features and cognitive variables. In this context, deep-learning classifiers have recently gained traction as a leading approach to detect cognitive variables from measured features [315]. Yet, their limited interpretability hampers trust in inferences based on classification performance. In principle, DreaMR could help interrogate the specific spatiotemporal features of neural activity that contribute to classifier decisions, thereby enhancing the utility of classification analyses in neuroimaging studies.

Chapter 6

Conclusion

In this thesis, we introduced a series of innovative methodologies to enhance the classification performance and interpretability of fMRI time series analysis. First, BolT, a novel transformer model, was presented to improve classification performance on fMRI time series. BolT employs fused window attention to capture local interactions among temporally-overlapped time windows and hierarchically grows window overlap to capture global representations. By integrating token fusion and cross-window regularization, BolT effectively integrates latent representations across the time series. Demonstrations on gender and disease detection from resting-state fMRI, as well as task detection from task-based fMRI, underscored the model’s potential for sensitive and explainable analysis of multivariate fMRI data. BolT’s ability to detect neurological disorders with characteristic influences on fMRI activation patterns and classify intricate task variables during cognitive processing further highlights its applicability.

Next, GraphCorr, a novel plug-in graph neural network, was introduced to enhance the performance of learning-based models for fMRI classification. GraphCorr employs node embedder and lag filter modules to extract dynamic and lagged functional connectivity features from whole-brain fMRI time series. This

transformation of raw BOLD signals into an efficient graph representation preserves fine-grained temporal information, which is often diminished in conventional functional connectivity features. By augmenting baseline classification models, GraphCorr significantly improves both performance and interpretability, making it a valuable tool for the analysis of fMRI time series.

Finally, DreaMR, a novel counterfactual explanation method for fMRI, was introduced based on a fractional multi-phase-distilled diffusion prior. DreaMR demonstrated superior sampling efficiency and fidelity compared to competing counterfactual methods, as evidenced by experiments on both resting-state and task-based fMRI. This method not only facilitates the interpretation of deep learning model decisions but also holds great promise for enabling explainable analysis of multivariate fMRI data.

Collectively, the approaches presented in this thesis contribute to the advancement of sensitive, efficient, and explainable fMRI data analysis, offering new possibilities for understanding complex brain functions and improving diagnostic tools for neurological disorders.

Bibliography

- [1] H. A. Bedel, I. Sivgin, O. Dalmaz, S. U. Dar, and T. Çukur, “BolT: Fused window transformers for fMRI time series analysis,” *Med Image Anal*, vol. 88, p. 102841, 2023.
- [2] B. Quist, B. A. Hargreaves, T. Cukur, G. R. Morrell, G. E. Gold, and N. K. Bangerter, “Simultaneous fat suppression and band reduction with large-angle multiple-acquisition balanced steady-state free precession,” *Magnetic resonance in medicine*, vol. 67, no. 4, pp. 1004–1012, 2012.
- [3] S. Ilbey, C. B. Top, A. Güngör, T. Çukur, E. U. Saritas, and H. E. Güven, “Fast system calibration with coded calibration scenes for magnetic particle imaging,” *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2070–2080, 2019.
- [4] E. Bıyık, K. Keskin, S. UH Dar, A. Koç, and T. Çukur, “Factorized sensitivity estimation for artifact suppression in phase-cycled bssfp mri,” *NMR in Biomedicine*, vol. 33, no. 4, p. e4228, 2020.
- [5] O. B. Demirel, T. Kilic, T. Çukur, and E. U. Saritas, “Anatomical measurements correlate with individual magnetostimulation thresholds for khz-range homogeneous magnetic fields,” *Medical physics*, vol. 47, no. 4, pp. 1836–1844, 2020.
- [6] B. Gözcü, R. K. Mahabadi, Y.-H. Li, E. Ilıcak, T. Cukur, J. Scarlett, and V. Cevher, “Learning-based compressive mri,” *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1394–1406, 2018.

- [7] T. Çukur and E. Ü. Sarıtaş, “Comparison of system-matrix-based and projection-based reconstructions for field free line magnetic particle imaging,” *International Journal on Magnetic Particle Imaging*, vol. 3, no. 1, pp. 1–8, 2017.
- [8] M. Aksoy, C. Forman, M. Straka, T. Çukur, J. Hornegger, and R. Bammer, “Hybrid prospective and retrospective head motion correction to mitigate cross-calibration errors,” *Magnetic resonance in medicine*, vol. 67, no. 5, pp. 1237–1251, 2012.
- [9] N. K. Bangerter, T. Cukur, B. A. Hargreaves, B. S. Hu, J. H. Brittain, D. Park, G. E. Gold, and D. G. Nishimura, “Three-dimensional fluid-suppressed t2-prep flow-independent peripheral angiography using balanced ssfp,” *Magnetic resonance imaging*, vol. 29, no. 8, pp. 1119–1124, 2011.
- [10] T. Çukur, A. Shimakawa, H. Yu, B. A. Hargreaves, B. S. Hu, D. G. Nishimura, and J. H. Brittain, “Magnetization-prepared ideal bssfp: a flow-independent technique for noncontrast-enhanced peripheral angiography,” *Journal of Magnetic Resonance Imaging*, vol. 33, no. 4, pp. 931–939, 2011.
- [11] T. Çukur and D. G. Nishimura, “Fat-water separation with alternating repetition time balanced ssfp,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 60, no. 2, pp. 479–484, 2008.
- [12] A. Güngör, B. Askin, D. A. Soydan, C. B. Top, E. U. Saritas, and T. Çukur, “Deq-mpi: A deep equilibrium reconstruction with learned consistency for magnetic particle imaging,” *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 321–334, 2023.
- [13] A. Zaid Alkilani, T. Çukur, and E. U. Saritas, “Fd-net: An unsupervised deep forward-distortion model for susceptibility artifact correction in epi,” *Magnetic Resonance in Medicine*, vol. 91, no. 1, pp. 280–296, 2024.
- [14] I. Aytakin, O. Dalmaz, K. Gonc, H. Ankishan, E. U. Saritas, U. Bagci, H. Celik, and T. Çukur, “Covid-19 detection from respiratory sounds with

- hierarchical spectrogram transformers,” *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 3, pp. 1273–1284, 2024.
- [15] Q. Zhang, T. Cukur, H. Greenspan, and G. Yang, “Editorial: Generative adversarial networks in cardiovascular research,” *Frontiers in Cardiovascular Medicine*, vol. 10, 2023.
- [16] B. Bilgic and T. Cukur, “Parallel imaging and reconstruction techniques,” in *Advances in Magnetic Resonance Technology and Applications*, vol. 10, pp. 139–159, Elsevier, 2023.
- [17] Ş. Öztürk, E. Çelik, and T. Çukur, “Content-based medical image retrieval with opponent class adaptive margin loss,” *Information Sciences*, vol. 637, p. 118938, 2023.
- [18] A. Güngör, B. Askin, D. A. Soydan, E. U. Saritas, C. B. Top, and T. Çukur, “Transms: Transformers for super-resolution calibration in magnetic particle imaging,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3562–3574, 2022.
- [19] Ş. Öztürk and T. Çukur, “Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4679–4690, 2022.
- [20] K. Keskin, U. Yılmaz, and T. Çukur, “Constrained ellipse fitting for efficient parameter mapping with phase-cycled bssfp mri,” *arXiv preprint arXiv:2106.03239*, 2021.
- [21] A. Ozaslan, A. Alacaoglu, O. Demirel, T. Çukur, and E. Saritas, “Fully automated gridding reconstruction for non-cartesian x-space magnetic particle imaging,” *Physics in Medicine & Biology*, vol. 64, no. 16, p. 165018, 2019.
- [22] A. Güngör, B. B. Bilecen, and T. Çukur, “Bayesian conditioned diffusion models for inverse problems,” *arXiv preprint arXiv:2406.09768*, 2024.

- [23] S. U. H. Dar, Ş. Öztürk, M. Özbey, K. K. Oguz, and T. Çukur, “Parallel-stream fusion of scan-specific and scan-general priors for learning deep mri reconstruction in low-data regimes,” *Computers in Biology and Medicine*, vol. 167, p. 107610, 2023.
- [24] M. U. Mirza, O. Dalmaz, H. A. Bedel, G. Elmas, Y. Korkmaz, A. Gungor, S. U. Dar, and T. Çukur, “Learning fourier-constrained diffusion bridges for mri reconstruction,” *arXiv preprint arXiv:2308.01096*, 2023.
- [25] O. Dalmaz, M. U. Mirza, G. Elmas, M. Ozbey, S. U. Dar, E. Ceyani, K. K. Oguz, S. Avestimehr, and T. Çukur, “One model to unite them all: Personalized federated learning of multi-contrast mri synthesis,” *Medical Image Analysis*, vol. 94, p. 103121, 2024.
- [26] A. Güngör, S. U. Dar, Ş. Öztürk, Y. Korkmaz, H. A. Bedel, G. Elmas, M. Ozbey, and T. Çukur, “Adaptive diffusion priors for accelerated mri reconstruction,” *Medical image analysis*, vol. 88, p. 102872, 2023.
- [27] G. Elmas, S. U. Dar, Y. Korkmaz, E. Ceyani, B. Susam, M. Ozbey, S. Avestimehr, and T. Çukur, “Federated learning of generative image priors for MRI reconstruction,” *IEEE Trans Med Imaging*, vol. 42, no. 7, pp. 1996–2009, 2022.
- [28] E. Ilicak, E. U. Saritas, and T. Cukur, “Automated parameter selection for accelerated mri reconstruction via low-rank modeling of local k-space neighborhoods,” *Zeitschrift für Medizinische Physik*, vol. 33, no. 2, pp. 203–219, 2023.
- [29] Y. Korkmaz, S. U. Dar, M. Yurt, M. Özbey, and T. Cukur, “Unsupervised MRI reconstruction via zero-shot learned adversarial transformers,” *IEEE Trans Med Imaging*, vol. 41, no. 7, pp. 1747–1763, 2022.
- [30] M. Yurt, M. Özbey, S. U. Dar, B. Tinaz, K. K. Oguz, and T. Çukur, “Progressively volumetrized deep generative models for data-efficient contextual learning of mr image recovery,” *Medical Image Analysis*, vol. 78, p. 102429, 2022.

- [31] S. U. Dar, M. Yurt, M. Shahdloo, M. E. Ildız, B. Tınaz, and T. Çukur, “Prior-guided image reconstruction for accelerated multi-contrast mri via generative adversarial networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1072–1087, 2020.
- [32] S. U. H. Dar, M. Özbey, A. B. Çath, and T. Çukur, “A transfer-learning approach for accelerated mri using deep neural networks,” *Magnetic resonance in medicine*, vol. 84, no. 2, pp. 663–685, 2020.
- [33] S. U. H. Dar, M. Yurt, and T. Çukur, “A few-shot learning approach for accelerated mri via fusion of data-driven and subject-driven priors,” *arXiv preprint arXiv:2103.07790*, 2021.
- [34] O. F. Atli, B. Kabas, F. Arslan, M. Yurt, O. Dalmaz, and T. Çukur, “I2i-mamba: Multi-modal medical image synthesis via selective state space modeling,” *arXiv preprint arXiv:2405.14022*, 2024.
- [35] V. A. Nezhad, G. Elmas, F. Arslan, B. Kabaş, and T. Çukur, “Generalizable deep mri reconstruction with cross-site data synthesis,” in *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2024.
- [36] M. U. Mirza, F. Arslan, and T. Çukur, “Super resolution mri via upscaling diffusion bridges,” in *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2024.
- [37] B. Kabas, F. Arslan, V. A. Nezhad, and T. Çukur, “Multi-contrast mr image synthesis with a brownian diffusion model,” in *2024 32nd Signal Processing and Communications Applications Conference (SIU)*, pp. 1–4, IEEE, 2024.
- [38] F. Arslan, B. Kabas, O. Dalmaz, M. Ozbey, and T. Çukur, “Self-consistent recursive diffusion bridge for medical image translation,” *arXiv preprint arXiv:2405.06789*, 2024.
- [39] M. Özbey, O. Dalmaz, S. U. H. Dar, H. A. Bedel, Öztürk, A. Güngör, and T. Çukur, “Unsupervised medical image translation with adversarial

- diffusion models,” *IEEE Transactions on Medical Imaging*, vol. 42, no. 12, pp. 3524–3539, 2023.
- [40] S. U. Dar, M. Yurt, L. Karacan, A. Erdem, E. Erdem, and T. Çukur, “Image synthesis in multi-contrast MRI with conditional generative adversarial networks,” *IEEE Trans Med Imaging*, vol. 38, no. 10, pp. 2375–2388, 2019.
- [41] M. Yurt, O. Dalmaz, S. Dar, M. Ozbey, B. Tmaz, K. Oguz, and T. Çukur, “Semi-Supervised Learning of MRI Synthesis without Fully-Sampled Ground Truths,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 12, pp. 3895–3906, 2022.
- [42] M. Yurt, S. U. Dar, A. Erdem, E. Erdem, K. K. Oguz, and T. Çukur, “mustGAN: multi-stream generative adversarial networks for MR image synthesis,” vol. 70, p. 101944, 2021.
- [43] M. Shahdloo, E. Çelik, and T. Çukur, “Biased competition in semantic representation during natural visual search,” *Neuroimage*, vol. 216, p. 116383, 2020.
- [44] S. Tokgoz, D. Aydogdu, B. Ilhan, Y. Sahin, N. Bariseri, B. M. Ozturkler, and T. Çukur, “Musical mirror-symmetrical movement tasks: comparison of rhythm versus melody-playing,” *NeuroReport*, vol. 31, no. 7, pp. 523–529, 2020.
- [45] T. Çukur, S. Nishimoto, A. G. Huth, and J. L. Gallant, “Attention during natural vision warps semantic representation across the human brain,” *Nat Neurosci*, vol. 16, no. 6, pp. 763–770, 2013.
- [46] I. Kiremitçi, Ö. Yilmaz, E. Çelik, M. Shahdloo, A. G. Huth, and T. Çukur, “Attentional modulation of hierarchical speech representations in a multitalker environment,” *Cerebral Cortex*, vol. 31, no. 11, pp. 4986–5005, 2021.
- [47] F. T. Yarman Vural, S. D. Newman, T. Çukur, and I. Önal Ertugrul, “Machine learning methods for human brain imaging,” 2023.

- [48] M. Shahdloo, E. Çelik, B. A. Urgan, J. L. Gallant, and T. Çukur, “Task-dependent warping of semantic representations during search for visual action categories,” *Journal of Neuroscience*, vol. 42, no. 35, pp. 6782–6799, 2022.
- [49] L. K. Şenel, F. Şahinuç, V. Yücesoy, H. Schütze, T. Çukur, and A. Koç, “Learning interpretable word embeddings via bidirectional alignment of dimensions with semantic concepts,” *Information Processing & Management*, vol. 59, no. 3, p. 102925, 2022.
- [50] E. Çelik, U. Keles, İbrahim Kiremitçi, J. L. Gallant, and T. Çukur, “Cortical networks of dynamic scene category representation in the human brain,” *Cortex*, vol. 143, pp. 127–147, 2021.
- [51] T. Zhang, J. S. Gao, T. Çukur, and J. L. Gallant, “Voxel-based state space modeling recovers task-related cognitive states in naturalistic fmri experiments,” *Frontiers in neuroscience*, vol. 14, p. 1103, 2021.
- [52] Ö. Yılmaz, E. Çelik, and T. Çukur, “Informed feature regularization in voxelwise modeling for naturalistic fmri experiments,” *European Journal of Neuroscience*, vol. 52, no. 5, pp. 3394–3410, 2020.
- [53] E. Çelik, S. U. H. Dar, Ö. Yılmaz, Ü. Keleş, and T. Çukur, “Spatially informed voxelwise modeling for naturalistic fmri experiments,” *NeuroImage*, vol. 186, pp. 741–757, 2019.
- [54] M. Shahdloo, E. Çelik, B. A. Urgan, J. L. Gallant, and T. Çukur, “Task-dependent warping of semantic representations during search for visual action categories,” *bioRxiv*, 2021.
- [55] T. Çukur, A. G. Huth, S. Nishimoto, and J. L. Gallant, “Functional subdomains within human ffa,” *Journal of Neuroscience*, vol. 33, no. 42, pp. 16748–16766, 2013.
- [56] I. Sivgin, H. A. Bedel, S. Ozturk, and T. Çukur, “A plug-in graph neural network to boost temporal sensitivity in fmri analysis,” *IEEE Journal of Biomedical and Health Informatics*, 2024.

- [57] E. M. Hillman, “Coupling mechanism and significance of the BOLD signal: a status report,” *Annual Review of Neuroscience*, vol. 37, pp. 161–181, 2014.
- [58] J. C. Rajapakse, F. Kruggel, J. M. Maisog, and D. Yves von Cramon, “Modeling hemodynamic response for analysis of functional MRI time-series,” *Human Brain Mapping*, vol. 6, no. 4, pp. 283–300, 1998.
- [59] M. Kubicki, R. W. McCarley, P. G. Nestor, T. Huh, R. Kikinis, M. E. Shenton, and C. G. Wible, “An fMRI study of semantic processing in men with schizophrenia,” *NeuroImage*, vol. 20, no. 4, pp. 1923–1933, 2003.
- [60] D. Wang, L. Shi, D. S. Yeung, P.-A. Heng, T.-T. Wong, and E. C. Tsang, “Support vector clustering for brain activation detection,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 572–579, Springer, 2005.
- [61] J. M. Papma, M. Smits, M. De Groot, F. U. Mattace Raso, A. van der Lugt, H. A. Vrooman, W. J. Niessen, P. J. Koudstaal, J. C. van Swieten, F. M. van der Veen, *et al.*, “The effect of hippocampal function, volume and connectivity on posterior cingulate cortex functioning during episodic memory fMRI in mild cognitive impairment,” *European Radiology*, vol. 27, no. 9, pp. 3716–3724, 2017.
- [62] A. Mensch, J. Mairal, D. Bzdok, B. Thirion, and G. Varoquaux, “Learning neural representations of human cognition across many fMRI studies,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [63] K. Li, L. Guo, J. Nie, G. Li, and T. Liu, “Review of methods for functional brain connectivity detection using fMRI,” *Comput Med Imaging Graph*, vol. 33, no. 2, pp. 131–139, 2009.
- [64] A. Venkataraman, K. R. Van Dijk, R. L. Buckner, and P. Golland, “Exploring functional connectivity in fMRI via clustering,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 441–444, IEEE, 2009.

- [65] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, “Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies,” *Current Biology*, vol. 21, no. 19, pp. 1641–1646, 2011.
- [66] O. Simon, F. Kherif, G. Flandin, J.-B. Poline, D. Riviere, J.-F. Mangin, D. Le Bihan, and S. Dehaene, “Automatized clustering and functional geometry of human parietofrontal networks for language, space, and number,” *NeuroImage*, vol. 23, no. 3, pp. 1192–1202, 2004.
- [67] B. P. Rogers, V. L. Morgan, A. T. Newton, and J. C. Gore, “Assessing functional connectivity in the human brain by fMRI,” *Magnetic Resonance Imaging*, vol. 25, no. 10, pp. 1347–1357, 2007.
- [68] C. Niu, A. D. Cohen, X. Wen, Z. Chen, P. Lin, X. Liu, B. H. Menze, B. Wiestler, Y. Wang, and M. Zhang, “Modeling motor task activation from resting-state fMRI using machine learning in individual subjects,” *Brain Imaging and Behavior*, vol. 15, no. 1, pp. 122–132, 2021.
- [69] B. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, *et al.*, “The organization of the human cerebral cortex estimated by intrinsic functional connectivity,” *Journal of Neurophysiology*, 2011.
- [70] K. R. Van Dijk, T. Hedden, A. Venkataraman, K. C. Evans, S. W. Lazar, and R. L. Buckner, “Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization,” *Journal of Neurophysiology*, vol. 103, no. 1, pp. 297–321, 2010.
- [71] Z. Hu and P. Shi, “Interregional functional connectivity via pattern synchrony,” in *International Conference on Control, Automation, Robotics and Vision*, pp. 1–6, IEEE, 2006.
- [72] M. Greicius, “Resting-state functional connectivity in neuropsychiatric disorders,” *Current Opinion in Neurology*, vol. 21, no. 4, pp. 424–430, 2008.
- [73] B. Lei, S. Yu, X. Zhao, A. F. Frangi, E.-L. Tan, A. Elazab, T. Wang, and S. Wang, “Diagnosis of early Alzheimer’s disease based on dynamic high

- order networks,” *Brain Imaging and Behavior*, vol. 15, no. 1, pp. 276–287, 2021.
- [74] A. Iraj, R. R. Benson, R. D. Welch, B. J. O’Neil, J. L. Woodard, S. I. Ayaz, A. Kulek, V. Mika, P. Medado, H. Soltanian-Zadeh, *et al.*, “Resting state functional connectivity in mild traumatic brain injury at the acute stage: independent component and seed-based analyses,” *Journal of Neurotrauma*, vol. 32, no. 14, pp. 1031–1045, 2015.
- [75] Y. Zhang, H. Zhang, X. Chen, S.-W. Lee, and D. Shen, “Hybrid high-order functional connectivity networks using resting-state functional MRI for mild cognitive impairment diagnosis,” *Sci Rep*, vol. 7, no. 1, pp. 1–15, 2017.
- [76] R. Kong, J. Li, C. Orban, M. R. Sabuncu, H. Liu, A. Schaefer, N. Sun, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, *et al.*, “Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion,” *Cerebral Cortex*, vol. 29, no. 6, pp. 2533–2551, 2019.
- [77] K. Rajpoot, A. Riaz, W. Majeed, and N. Rajpoot, “Functional connectivity alterations in epilepsy from resting-state functional MRI,” *PloS One*, vol. 10, no. 8, p. e0134944, 2015.
- [78] E. M. Müller-Oehring, D. Kwon, B. J. Nagel, E. V. Sullivan, W. Chu, T. Rohlfing, D. Prouty, B. N. Nichols, J.-B. Poline, S. F. Tapert, *et al.*, “Influences of age, sex, and moderate alcohol drinking on the intrinsic functional architecture of adolescent brains,” *Cerebral Cortex*, vol. 28, no. 3, pp. 1049–1063, 2018.
- [79] J. S. Anderson, J. A. Nielsen, M. A. Ferguson, M. C. Burbach, E. T. Cox, L. Dai, G. Gerig, J. O. Edgin, and J. R. Korenberg, “Abnormal brain synchrony in Down syndrome,” *NeuroImage: Clinical*, vol. 2, pp. 703–715, 2013.
- [80] M. J. McKeown and T. J. Sejnowski, “Independent component analysis of fMRI data: examining the assumptions,” *Human Brain Mapping*, vol. 6, no. 5-6, pp. 368–372, 1998.

- [81] M. Svensén, F. Kruggel, and H. Benali, “ICA of fMRI group study data,” *NeuroImage*, vol. 16, no. 3, pp. 551–563, 2002.
- [82] F. Pereira, T. Mitchell, and M. Botvinick, “Machine learning classifiers and fMRI: a tutorial overview,” *Neuroimage*, vol. 45, no. 1, pp. S199–S209, 2009.
- [83] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, “Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns,” *NeuroImage*, vol. 43, no. 1, pp. 44–58, 2008.
- [84] X. Zhang, B. Hu, X. Ma, and L. Xu, “Resting-state whole-brain functional connectivity networks for MCI classification using L2-regularized logistic regression,” *IEEE Transactions on Nanobioscience*, vol. 14, no. 2, pp. 237–247, 2015.
- [85] C. Wang, Z. Xiao, and J. Wu, “Functional connectivity-based classification of autism and control using SVM-RFECV on rs-fMRI data,” *Physica Medica*, vol. 65, pp. 99–105, 2019.
- [86] A. S. Heinsfeld, A. R. Franco, R. C. Craddock, A. Buchweitz, and F. Meneguzzi, “Identification of autism spectrum disorder using deep learning and the ABIDE dataset,” *NeuroImage: Clin*, vol. 17, pp. 16–23, 2018.
- [87] Y. Li, J. Liu, Z. Tang, and B. Lei, “Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for MCI identification,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 9, pp. 2818–2830, 2020.
- [88] J. S. Duncan, M. F. Insana, and N. Ayache, “Biomedical imaging and analysis in the age of big data and deep learning [scanning the issue],” *Proceedings of the IEEE*, vol. 108, no. 1, pp. 3–10, 2019.
- [89] P. Mlynarski, H. Delingette, A. Criminisi, and N. Ayache, “Deep learning with mixed supervision for brain tumor segmentation,” *Journal of Medical Imaging*, vol. 6, no. 3, p. 034002, 2019.

- [90] T.-E. Kam, H. Zhang, Z. Jiao, and D. Shen, “Deep learning of static and dynamic brain functional networks for early MCI detection,” *IEEE Trans Med Imaging*, vol. 39, no. 2, pp. 478–487, 2019.
- [91] J. Kawahara, C. J. Brown, S. P. Miller, B. G. Booth, V. Chau, R. E. Grunau, J. G. Zwicker, and G. Hamarneh, “BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment,” *NeuroImage*, vol. 146, pp. 1038–1049, 2017.
- [92] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. Guerrero, B. Glocker, and D. Rueckert, “Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer’s disease,” *Medical Image Analysis*, vol. 48, pp. 117–130, 2018.
- [93] L. Fan, J. Su, J. Qin, D. Hu, and H. Shen, “A deep network model on dynamic functional connectivity with applications to gender classification and intelligence prediction,” *Frontiers in Neuroscience*, vol. 14, p. 881, 2020.
- [94] L. Wang, K. Li, and X. P. Hu, “Graph convolutional network for fMRI analysis based on connectivity neighborhood,” *Network Neuroscience*, vol. 5, no. 1, pp. 83–95, 2021.
- [95] P.-J. Lahaye, J.-B. Poline, G. Flandin, S. Dodel, and L. Garnero, “Functional connectivity: studying nonlinear, delayed interactions between BOLD signals,” *NeuroImage*, vol. 20, no. 2, pp. 962–974, 2003.
- [96] Z. Hu and P. Shi, “Nonlinear analysis of BOLD signal: biophysical modeling, physiological states, and functional activation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 734–741, Springer, 2007.
- [97] N. C. Dvornek, P. Ventola, K. A. Pelphrey, and J. S. Duncan, “Identifying autism from resting-state fMRI using long short-term memory networks,” in *Int Work Mach Learn Med Imaging*, pp. 362–370, 2017.
- [98] S. Nguyen, B. Ng, A. D. Kaplan, and P. Ray, “Attend and decode: 4d fMRI task state decoding using attention models,” in *Mach Learn Health*, pp. 267–279, PMLR, 2020.

- [99] I. Malkiel, G. Rosenman, L. Wolf, and T. Hendler, “Pre-training and fine-tuning transformers for fMRI prediction tasks,” *arXiv:2112.05761*, 2021.
- [100] A. A. Ismail, M. Gunady, L. Pessoa, H. Corrada Bravo, and S. Feizi, “Input-cell attention reduces vanishing saliency of recurrent neural networks,” *Adv Neural Inf Process Syst*, vol. 32, 2019.
- [101] R. Liégeois, J. Li, R. Kong, C. Orban, D. Van De Ville, T. Ge, M. R. Sabuncu, and B. Yeo, “Resting brain dynamics at different timescales capture distinct aspects of human behavior,” *Nature Communications*, vol. 10, no. 1, pp. 1–9, 2019.
- [102] E. A. Allen, E. Damaraju, S. M. Plis, E. B. Erhardt, T. Eichele, and V. D. Calhoun, “Tracking whole-brain connectivity dynamics in the resting state,” *Cereb Cortex*, vol. 24, no. 3, pp. 663–676, 2014.
- [103] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, *et al.*, “The wu-minn human connectome project: an overview,” *NeuroImage*, vol. 80, pp. 62–79, 2013.
- [104] A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, *et al.*, “The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism,” *Molecular psychiatry*, vol. 19, no. 6, pp. 659–667, 2014.
- [105] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.
- [106] M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith, “Temporal autocorrelation in univariate linear modeling of fMRI data,” *NeuroImage*, vol. 14, no. 6, pp. 1370–1386, 2001.
- [107] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby, “Beyond mind-reading: multi-voxel pattern analysis of fMRI data,” *Trend Cognit Sci*, vol. 10, no. 9, pp. 424–430, 2006.

- [108] J. V. Haxby, “Multivariate pattern analysis of fMRI: the early beginnings,” *NeuroImage*, vol. 62, no. 2, pp. 852–855, 2012.
- [109] X. Song and N.-k. Chen, “A SVM-based quantitative fMRI method for resting-state functional network detection,” *Magnetic Resonance Imaging*, vol. 32, no. 7, pp. 819–831, 2014.
- [110] Z. Wang, A. R. Childress, J. Wang, and J. A. Detre, “Support vector machine learning-based fMRI data group analysis,” *NeuroImage*, vol. 36, no. 4, pp. 1139–1151, 2007.
- [111] S. H. Hojjati, A. Ebrahimzadeh, A. Khazaei, A. Babajani-Feremi, A. D. N. Initiative, *et al.*, “Predicting conversion from MCI to AD using resting-state fMRI, graph theoretical approach and SVM,” *Journal of Neuroscience Methods*, vol. 282, pp. 69–80, 2017.
- [112] E. Bullmore, S. Rabe-Hesketh, R. Morris, S. Williams, L. Gregory, J. Gray, and M. Brammer, “Functional magnetic resonance image analysis of a large-scale neurocognitive network,” *NeuroImage*, vol. 4, no. 1, pp. 16–33, 1996.
- [113] S.-y. Xie, R. Guo, N.-f. Li, G. Wang, and H.-t. Zhao, “Brain fMRI processing and classification based on combination of PCA and SVM,” in *Int Joint Conf Neural Netw*, pp. 3384–3389, IEEE, 2009.
- [114] R. A. Poldrack, “Region of interest analysis for fMRI,” *Social cognitive and affective neuroscience*, vol. 2, no. 1, pp. 67–70, 2007.
- [115] L.-L. Zeng, H. Shen, L. Liu, L. Wang, B. Li, P. Fang, Z. Zhou, Y. Li, and D. Hu, “Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis,” *Brain*, vol. 135, no. 5, pp. 1498–1507, 2012.
- [116] H. Shen, L. Wang, Y. Liu, and D. Hu, “Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI,” *NeuroImage*, vol. 49, no. 4, pp. 3110–3121, 2010.

- [117] A. Khazaee, A. Ebrahimzadeh, and A. Babajani-Feremi, “Application of advanced machine learning methods on resting-state fMRI network for identification of mild cognitive impairment and Alzheimer’s disease,” *Brain Imaging and Behavior*, vol. 10, no. 3, pp. 799–817, 2016.
- [118] H.-I. Suk, C.-Y. Wee, S.-W. Lee, and D. Shen, “State-space model with deep learning for functional dynamics estimation in resting-state fMRI,” *NeuroImage*, vol. 129, pp. 292–307, 2016.
- [119] S. Koyamada, Y. Shikauchi, K. Nakae, M. Koyama, and S. Ishii, “Deep learning of fMRI big data: a novel approach to subject-transfer decoding,” *arXiv:1502.00093*, 2015.
- [120] H. Huang, X. Hu, Y. Zhao, M. Makkie, Q. Dong, S. Zhao, L. Guo, and T. Liu, “Modeling task fMRI data via deep convolutional autoencoder,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1551–1561, 2017.
- [121] S. Sarraf and G. Tofghi, “Classification of Alzheimer’s disease using fMRI data and deep learning convolutional neural networks,” *arXiv:1603.08631*, 2016.
- [122] S. Sarraf and G. Tofghi, “Deep learning-based pipeline to recognize Alzheimer’s disease using fMRI data,” in *Future Technologies Conference*, pp. 816–820, IEEE, 2016.
- [123] Y. Zhao, Q. Dong, S. Zhang, W. Zhang, H. Chen, X. Jiang, L. Guo, X. Hu, J. Han, and T. Liu, “Automatic recognition of fMRI-derived functional networks using 3-D convolutional neural networks,” *IEEE Trans Biomed Eng*, vol. 65, no. 9, pp. 1975–1984, 2017.
- [124] R. J. Meszlényi, K. Buza, and Z. Vidnyánszky, “Resting state fMRI functional connectivity-based classification using a convolutional neural network architecture,” *Front Neuroinf*, vol. 11, p. 61, 2017.
- [125] X. Xing, Q. Li, H. Wei, M. Zhang, Y. Zhan, X. S. Zhou, Z. Xue, and F. Shi, “Dynamic spectral graph convolution networks with assistant task

- training for early mci diagnosis,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 639–646, Springer, 2019.
- [126] X. Li, Y. Zhou, N. Dvornek, M. Zhang, S. Gao, J. Zhuang, D. Scheinost, L. H. Staib, P. Ventola, and J. S. Duncan, “Braingnn: Interpretable brain graph neural network for fMRI analysis,” *Med Image Anal*, vol. 74, p. 102233, 2021.
 - [127] X. Li, N. C. Dvornek, Y. Zhou, J. Zhuang, P. Ventola, and J. S. Duncan, “Graph neural network for interpreting task-fMRI biomarkers,” in *Int Conf Med Image Comput Comput Assist Interv*, pp. 485–493, 2019.
 - [128] B.-H. Kim, J. C. Ye, and J.-J. Kim, “Learning dynamic graph representation of brain connectome with spatio-temporal attention,” *Adv Neural Inf Process Syst*, vol. 34, 2021.
 - [129] D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun, “Measuring and relieving the over-smoothing problem for graph neural networks from the topological view,” in *AAAI Conference on Artificial Intelligence*, vol. 34, pp. 3438–3445, 2020.
 - [130] U. Alon and E. Yahav, “On the bottleneck of graph neural networks and its practical implications,” *arXiv:2006.05205*, 2020.
 - [131] L. Su, L. Wang, H. Shen, G. Feng, and D. Hu, “Discriminative analysis of non-linear brain connectivity in schizophrenia: an fMRI Study,” *Frontiers in Human Neuroscience*, vol. 7, p. 702, 2013.
 - [132] W. Li, X. Lin, and X. Chen, “Detecting Alzheimer’s disease Based on 4D fMRI: An exploration under deep learning framework,” *Neurocomputing*, vol. 388, pp. 280–287, 2020.
 - [133] C. Zhao, H. Li, Z. Jiao, T. Du, and Y. Fan, “A 3d convolutional encapsulated long short-term memory (3dconv-lstm) model for denoising fmri data,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 479–488, Springer, 2020.

- [134] G. Kerg, B. Kanuparthi, A. G. ALIAS PARTH GOYAL, K. Goyette, Y. Bengio, and G. Lajoie, “Untangling tradeoffs between recurrence and self-attention in artificial neural networks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 19443–19454, 2020.
- [135] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [136] J. Zhang, L. Zhou, L. Wang, M. Liu, and D. Shen, “Diffusion Kernel Attention Network for Brain Disorder Classification,” *IEEE Transactions on Medical Imaging*, 2022.
- [137] A. Abraham, M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux, “Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example,” *NeuroImage*, vol. 147, pp. 736–745, 2017.
- [138] S. Parisot, S. I. Ktena, E. Ferrante, M. Lee, R. G. Moreno, B. Glocker, and D. Rueckert, “Spectral graph convolutions for population-based disease prediction,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 177–185, Springer, 2017.
- [139] S. Gadgil, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, E. Adeli, and K. M. Pohl, “Spatio-temporal graph convolution for resting-state fMRI analysis,” in *Int Conf Med Image Comput Comput Assist Interv*, pp. 528–538, Springer, 2020.
- [140] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.
- [141] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The long-document transformer,” *arXiv preprint arXiv:2004.05150*, 2020.

- [142] C. Zhao, W. Chen, J. Qin, P. Yang, Z. Xiang, A. F. Frangi, M. Chen, S. Fan, W. Yu, X. Chen, *et al.*, “Ift-net: Interactive fusion transformer network for quantitative analysis of pediatric echocardiography,” *Medical Image Analysis*, vol. 82, p. 102648, 2022.
- [143] S. Mehta, X. Lu, W. Wu, D. Weaver, H. Hajishirzi, J. G. Elmore, and L. G. Shapiro, “End-to-end diagnosis of breast biopsy images with transformers,” *Medical Image Analysis*, vol. 79, p. 102466, 2022.
- [144] R. M. Hutchison, T. Womelsdorf, E. A. Allen, P. A. Bandettini, V. D. Calhoun, M. Corbetta, S. Della Penna, J. H. Duyn, G. H. Glover, J. Gonzalez-Castillo, *et al.*, “Dynamic functional connectivity: promise, issues, and interpretations,” *NeuroImage*, vol. 80, pp. 360–378, 2013.
- [145] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv:2010.11929*, 2020.
- [146] J. Yang, C. Li, P. Zhang, X. Dai, B. Xiao, L. Yuan, and J. Gao, “Focal self-attention for local-global interactions in vision transformers,” *arXiv:2107.00641*, 2021.
- [147] H. Chefer, S. Gur, and L. Wolf, “Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers,” in *IEEE/CVF International Conference on Computer Vision*, pp. 397–406, 2021.
- [148] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, *et al.*, “The minimal preprocessing pipelines for the Human Connectome Project,” *NeuroImage*, vol. 80, pp. 105–124, 2013.
- [149] C. Craddock, Y. Benhajali, C. Chu, F. Chouinard, A. Evans, A. Jakab, B. S. Khundrakpam, J. D. Lewis, Q. Li, M. Milham, *et al.*, “The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives,” *Frontiers in Neuroinformatics*, vol. 7, 2013.

- [150] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. Yeo, “Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI,” *Cereb Cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.
- [151] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, “Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni MRI single-subject brain,” *NeuroImage*, vol. 15, no. 1, pp. 273–289, 2002.
- [152] S. G. Kafali, T. Çukur, and E. U. Saritas, “Phase-correcting non-local means filtering for diffusion-weighted imaging of the spinal cord,” *Magnetic resonance in medicine*, vol. 80, no. 3, pp. 1020–1035, 2018.
- [153] E. Tagliazucchi, P. Balenzuela, D. Fraiman, and D. R. Chialvo, “Criticality in large-scale brain fMRI dynamics unveiled by a novel point process analysis,” *Front Physiol*, vol. 3, p. 15, 2012.
- [154] E. Tagliazucchi, P. Balenzuela, D. Fraiman, P. Montoya, and D. R. Chialvo, “Spontaneous BOLD event triggered averages for estimating functional connectivity at resting state,” *Neurosci Lett*, vol. 488, no. 2, pp. 158–163, 2011.
- [155] M. Rahman, U. Mahmood, N. Lewis, H. Gazula, A. Fedorov, Z. Fu, V. D. Calhoun, S. M. Plis, *et al.*, “Interpreting models interpreting brain dynamics,” *Sci Rep*, vol. 12, no. 1, pp. 1–15, 2022.
- [156] S. J. Ritchie, S. R. Cox, X. Shen, M. V. Lombardo, L. M. Reus, C. Alloza, M. A. Harris, H. L. Alderson, S. Hunter, E. Neilson, *et al.*, “Sex differences in the adult human brain: evidence from 5216 UK biobank participants,” *Cereb Cortex*, vol. 28, no. 8, pp. 2959–2975, 2018.
- [157] M. Filippi, P. Valsasina, P. Misci, A. Falini, G. Comi, and M. A. Rocca, “The organization of intrinsic brain activity differs between genders: A resting-state fMRI study in a large cohort of young healthy subjects,” *Hum Brain Mapp*, vol. 34, no. 6, pp. 1330–1343, 2013.

- [158] E. U. Saritas, D. Lee, T. Cukur, A. Shankaranarayanan, and D. G. Nishimura, “Hadamard slice encoding for reduced-fov diffusion-weighted imaging,” *Magnetic resonance in Medicine*, vol. 72, no. 5, pp. 1277–1290, 2014.
- [159] L. Q. Uddin, B. Yeo, and R. N. Spreng, “Towards a universal taxonomy of macro-scale functional human brain networks,” *Brain Topography*, vol. 32, no. 6, pp. 926–942, 2019.
- [160] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter, “The brain’s default network: anatomy, function, and relevance to disease,” *Annals of the New York Academy of Sciences*, vol. 1124, no. 1, pp. 1–38, 2008.
- [161] Y.-Y. Chen, M. Uljarevic, J. Neal, S. Greening, H. Yim, and T.-H. Lee, “Excessive functional coupling with less variability between salience and default-mode networks in Autism Spectrum Disorder,” *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2021.
- [162] R. VanRullen and L. Reddy, “Reconstructing faces from fMRI patterns using deep generative neural networks,” *Communications Biology*, vol. 2, no. 1, pp. 1–10, 2019.
- [163] H. Li, T. D. Satterthwaite, and Y. Fan, “Brain age prediction based on resting-state functional connectivity patterns using convolutional neural networks,” in *IEEE Int Symp Biomed Imaging*, pp. 101–104, IEEE, 2018.
- [164] S. M. LaConte, “Decoding fMRI brain states in real-time,” *NeuroImage*, vol. 56, no. 2, pp. 440–454, 2011.
- [165] P. Andersson, J. P. Pluim, J. C. Siero, S. Klein, M. A. Viergever, and N. F. Ramsey, “Real-time decoding of brain responses to visuospatial attention using 7T fMRI,” *PloS One*, vol. 6, no. 11, p. e27638, 2011.
- [166] S. Nishimoto, “Modeling movie-evoked human brain activity using motion-energy and space-time vision transformer features,” *bioRxiv*, 2021.

- [167] A. J. Anderson, B. D. Zinszer, and R. D. Raizada, “Representational similarity encoding for fMRI: Pattern-based synthesis to predict brain activity using stimulus-model-similarities,” *NeuroImage*, vol. 128, pp. 44–53, 2016.
- [168] G. H. Ngo, M. Nguyen, N. F. Chen, and M. R. Sabuncu, “A transformer-Based neural language model that synthesizes brain activation maps from free-form text queries,” *Medical Image Analysis*, vol. 81, p. 102540, 2022.
- [169] L. Q. Uddin, K. Supekar, and V. Menon, “Typical and atypical development of functional human brain networks: insights from resting-state fMRI,” *Frontiers in Systems Neuroscience*, p. 21, 2010.
- [170] C. Hohenfeld, C. J. Werner, and K. Reetz, “Resting-state connectivity in neurodegenerative disorders: Is there potential for an imaging biomarker?,” *NeuroImage: Clinical*, vol. 18, pp. 849–870, 2018.
- [171] N. C. Dvornek, P. Ventola, and J. S. Duncan, “Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks,” in *IEEE International Symposium on Biomedical Imaging*, pp. 725–728, IEEE, 2018.
- [172] G. Román and B. Pascual, “Contribution of neuroimaging to the diagnosis of Alzheimer’s disease and vascular dementia,” *Archives of Medical Research*, vol. 43, no. 8, pp. 671–676, 2012.
- [173] G. Flandin, F. Kherif, X. Pennec, G. Malandain, N. Ayache, and J.-B. Poline, “Improved detection sensitivity in functional MRI data using a brain parcelling technique,” in *Int Conf Med Image Comput Comput Assist Interv*, pp. 467–474, Springer, 2002.
- [174] M. Shahdloo, E. Ilıcak, M. Tofghi, E. U. Saritas, A. E. Çetin, and T. Çukur, “Projection onto epigraph sets for rapid self-tuning compressed sensing mri,” *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1677–1689, 2018.
- [175] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv:1810.04805*, 2018.

- [176] O. Dalmaz, M. Yurt, and T. Çukur, “ResViT: residual vision transformers for multi-modal medical image synthesis,” *arXiv preprint arXiv:2106.16031*, 2021.
- [177] A. Güngör, S. U. Dar, Ş. Öztürk, Y. Korkmaz, G. Elmas, M. Özbey, and T. Çukur, “Adaptive diffusion priors for accelerated MRI reconstruction,” *Med Image Anal*, vol. 88, p. 102872, 2023.
- [178] A. Güngör, E. Kopanoğlu, T. Çukur, and H. E. Güven, “Fast recovery of compressed multi-contrast magnetic resonance images,” in *Medical Imaging 2017: Image Processing*, vol. 10133, pp. 503–508, SPIE, 2017.
- [179] K. T. Kwon, H. H. Wu, T. Shin, T. Çukur, M. Lustig, and D. G. Nishimura, “Three-dimensional magnetization-prepared imaging using a concentric cylinders trajectory,” *Magnetic resonance in medicine*, vol. 71, no. 5, pp. 1700–1710, 2014.
- [180] G. Lanza, R. Bella, S. Giuffrida, M. Cantone, G. Pennisi, C. Spampinato, D. Giordano, G. Malaguarnera, A. Raggi, and M. Pennisi, “Preserved transcallosal inhibition to transcranial magnetic stimulation in nondemented elderly patients with leukoaraiosis,” *BioMed Res Int*, vol. 2013, 2013.
- [181] R. Bella, R. Ferri, M. Pennisi, M. Cantone, G. Lanza, G. Malaguarnera, C. Spampinato, D. Giordano, G. Alagona, and G. Pennisi, “Enhanced motor cortex facilitation in patients with vascular cognitive impairment-no dementia,” *Neurosci Lett*, vol. 503, no. 3, pp. 171–175, 2011.
- [182] D. Zhu, T. Zhang, X. Jiang, X. Hu, H. Chen, N. Yang, J. Lv, J. Han, L. Guo, and T. Liu, “Fusing DTI and fMRI data: a survey of methods and applications,” *NeuroImage*, vol. 102, pp. 184–191, 2014.
- [183] E. Bullmore and O. Sporns, “Complex brain networks: graph theoretical analysis of structural and functional systems,” *Nat Rev Neurosci*, vol. 10, no. 3, pp. 186–198, 2009.
- [184] E. E. Tripoliti, D. I. Fotiadis, and M. Argyropoulou, “A supervised method to assist the diagnosis and monitor progression of Alzheimer’s disease using

- data from an fMRI experiment,” *Artif Intell Med*, vol. 53, no. 1, pp. 35–45, 2011.
- [185] J. Pan, B. Lei, Y. Shen, Y. Liu, Z. Feng, and S. Wang, “Characterization multimodal connectivity of brain network by hypergraph GAN for Alzheimer’s disease analysis,” in *Pattern Recognit Comput Vis*, pp. 467–478, Springer, 2021.
 - [186] J. Mourao-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, “Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data,” *NeuroImage*, vol. 28, no. 4, pp. 980–995, 2005.
 - [187] B. Rashid, M. R. Arbabshirani, E. Damaraju, M. S. Cetin, R. Miller, G. D. Pearlson, and V. D. Calhoun, “Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity,” *NeuroImage*, vol. 134, pp. 645–657, 2016.
 - [188] N. U. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, *et al.*, “Prediction of individual brain maturity using fMRI,” *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010.
 - [189] M. D. Rosenberg, E. S. Finn, D. Scheinost, X. Papademetris, X. Shen, R. T. Constable, and M. M. Chun, “A neuromarker of sustained attention from whole-brain functional connectivity,” *Nat Neurosci*, vol. 19, no. 1, pp. 165–171, 2016.
 - [190] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. A. Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen, J. A. Turner, and V. D. Calhoun, “Deep learning for neuroimaging: a validation study,” *Front Neurosci*, vol. 8, p. 229, 2014.
 - [191] J. Kim, V. D. Calhoun, E. Shim, and J.-H. Lee, “Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain

- resting-state functional connectivity patterns of schizophrenia,” *NeuroImage*, vol. 124, pp. 127–146, 2016.
- [192] T. Eslami and F. Saeed, “Auto-ASD-network: a technique based on deep learning and support vector machines for diagnosing autism spectrum disorder using fMRI data,” in *Proc ACM Int Conf Bioinf Comput Biol Health Inf*, pp. 646–651, 2019.
- [193] R. D. Hjelm, V. D. Calhoun, R. Salakhutdinov, E. A. Allen, T. Adali, and S. M. Plis, “Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks,” *NeuroImage*, vol. 96, pp. 245–260, 2014.
- [194] S. Ilbey, C. B. Top, T. Çukur, E. U. Saritas, and H. E. Güven, “Image reconstruction for magnetic particle imaging using an augmented lagrangian method,” in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pp. 64–67, IEEE, 2017.
- [195] L. Shao, C. Fu, Y. You, and D. Fu, “Classification of ASD based on fMRI data with deep learning,” *Cognit Neurodyn*, vol. 15, no. 6, pp. 961–974, 2021.
- [196] M. Saeidi, W. Karwowski, F. V. Farahani, K. Fiok, P. Hancock, B. D. Sawyer, L. Christov-Moore, and P. K. Douglas, “Decoding task-based fMRI data with graph neural networks, considering individual differences,” *Brain Sci*, vol. 12, no. 8, p. 1094, 2022.
- [197] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *arXiv:1609.02907*, 2016.
- [198] G. Qu, W. Hu, L. Xiao, J. Wang, Y. Bai, B. Patel, K. Zhang, and Y.-P. Wang, “Brain functional connectivity analysis via graphical deep learning,” *IEEE Trans Biomed Eng*, vol. 69, no. 5, pp. 1696–1706, 2021.
- [199] A. Bessadok, M. A. Mahjoub, and I. Rekik, “Graph neural networks in network neuroscience,” *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 5, pp. 5833–5848, 2022.

- [200] X. Yu, L. Zhang, L. Zhao, Y. Lyu, T. Liu, and D. Zhu, “Disentangling spatial-temporal functional brain networks via twin-transformers,” *arXiv:2204.09225*, 2022.
- [201] W. Dai, Z. Zhang, L. Tian, S. Yu, S. Wang, Z. Dong, and H. Zheng, “Brain-former: A hybrid cnn-transformer model for brain fMRI data classification,” *arXiv:2208.03028*, 2022.
- [202] E. T. Rolls, M. Joliot, and N. Tzourio-Mazoyer, “Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas,” *NeuroImage*, vol. 122, pp. 1–5, 2015.
- [203] A. R. Laird, S. B. Eickhoff, F. Kurth, P. M. Fox, A. M. Uecker, J. A. Turner, J. L. Robinson, J. L. Lancaster, and P. T. Fox, “Ale meta-analysis workflows via the brainmap database: progress towards a probabilistic functional brain atlas,” *Front Neuroinf*, p. 23, 2009.
- [204] B.-H. Kim and J. C. Ye, “Understanding graph isomorphism network for rs-fMRI functional connectivity analysis,” *Front Neurosci*, p. 630, 2020.
- [205] T. Çukur, J. M. Santos, D. G. Nishimura, and J. M. Pauly, “Varying kernel-extent gridding reconstruction for undersampled variable-density spirals,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 59, no. 1, pp. 196–201, 2008.
- [206] J. Gan, Z. Peng, X. Zhu, R. Hu, J. Ma, and G. Wu, “Brain functional connectivity analysis based on multi-graph fusion,” *Med Image Anal*, vol. 71, p. 102057, 2021.
- [207] A. D. Savva, G. D. Mitsis, and G. K. Matsopoulos, “Assessment of dynamic functional connectivity in resting-state fMRI using the sliding window technique,” *Brain Behavior*, vol. 9, no. 4, p. e01255, 2019.
- [208] D. A. Handwerker, V. Roopchansingh, J. Gonzalez-Castillo, and P. A. Bandettini, “Periodic changes in fMRI connectivity,” *NeuroImage*, vol. 63, no. 3, pp. 1712–1719, 2012.

- [209] C. Zhang, C. C. Dougherty, S. A. Baum, T. White, and A. M. Michael, “Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity,” *Hum Brain Mapp*, vol. 39, no. 4, pp. 1765–1776, 2018.
- [210] L. Snoek, M. M. van der Miesen, T. Beemsterboer, A. van der Leij, A. Eigenhuis, and H. Steven Scholte, “The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses,” *Sci Data*, vol. 8, no. 1, pp. 1–23, 2021.
- [211] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” *Adv Neural Inf Process Syst*, vol. 30, 2017.
- [212] D. Z. Bolling, N. B. Pitskel, B. Deen, M. J. Crowley, J. C. McPartland, M. D. Kaiser, B. C. Vander Wyk, J. Wu, L. C. Mayes, and K. A. Pelphrey, “Enhanced neural responses to rule violation in children with autism: a comparison to social exclusion,” *Develop Cognit Neurosci*, vol. 1, no. 3, pp. 280–294, 2011.
- [213] T. Çukur, J. M. Santos, J. M. Pauly, and D. G. Nishimura, “Variable-density parallel imaging with partially localized coil sensitivities,” *IEEE transactions on medical imaging*, vol. 29, no. 5, pp. 1173–1181, 2010.
- [214] E. Biyik, E. Ilicak, and T. Cukur, “Reconstruction by calibration over tensors for multi-coil multi-acquisition balanced ssfp imaging,” *Magnetic resonance in medicine*, vol. 79, no. 5, pp. 2542–2554, 2018.
- [215] C.-Y. Wee, P.-T. Yap, D. Zhang, K. Denny, J. N. Browndyke, G. G. Potter, K. A. Welsh-Bohmer, L. Wang, and D. Shen, “Identification of MCI individuals using structural and functional connectivity networks,” *NeuroImage*, vol. 59, no. 3, pp. 2045–2056, 2012.
- [216] S. M. Smith, D. Vidaurre, C. F. Beckmann, M. F. Glasser, M. Jenkinson, K. L. Miller, T. E. Nichols, E. C. Robinson, G. Salimi-Khorshidi, M. W. Woolrich, *et al.*, “Functional connectomics from resting-state fMRI,” *Trend Cognit Sci*, vol. 17, no. 12, pp. 666–682, 2013.

- [217] F. Zhao, H. Zhang, I. Rekik, Z. An, and D. Shen, “Diagnosis of autism spectrum disorders using multi-level high-order functional networks derived from resting-state functional MRI,” *Front Hum Neurosci*, vol. 12, p. 184, 2018.
- [218] A. Riaz, M. Asad, E. Alonso, and G. Slabaugh, “DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI,” *J Neurosci Methods*, vol. 335, p. 108506, 2020.
- [219] L.-L. Zeng, H. Wang, P. Hu, B. Yang, W. Pu, H. Shen, X. Chen, Z. Liu, H. Yin, Q. Tan, *et al.*, “Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI,” *EBioMedicine*, vol. 30, pp. 74–85, 2018.
- [220] E. Ilıcak, L. K. Senel, E. Biyik, and T. Çukur, “Profile-encoding reconstruction for multiple-acquisition balanced steady-state free precession imaging,” *Magnetic resonance in medicine*, vol. 78, no. 4, pp. 1316–1329, 2017.
- [221] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, “A comprehensive survey on graph neural networks,” *IEEE Trans Neural Netw Learn Syst*, vol. 32, no. 1, pp. 4–24, 2020.
- [222] X. Liu, Y. Tian, Q. He, W.-C. Lee, and J. McPherson, “Distributed graph summarization,” in *Proc ACM Int Conf Inf Know Manage*, pp. 799–808, 2014.
- [223] D. Cai, X. He, J. Han, and T. S. Huang, “Graph regularized nonnegative matrix factorization for data representation,” *IEEE Trans Pattern Anal Mach Intell*, vol. 33, no. 8, pp. 1548–1560, 2010.
- [224] P. Han, P. Yang, P. Zhao, S. Shang, Y. Liu, J. Zhou, X. Gao, and P. Kalnis, “GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization,” in *Proc ACM SIGKDD*, pp. 705–713, 2019.
- [225] E. Kopanoglu, A. Güngör, T. Kilic, E. U. Saritas, K. K. Oguz, T. Çukur, and H. E. Güven, “Simultaneous use of individual and joint regularization

- terms in compressive sensing: Joint reconstruction of multi-channel multi-contrast mri acquisitions,” *NMR in Biomedicine*, vol. 33, no. 4, p. e4247, 2020.
- [226] Ü. Sakoğlu, G. D. Pearlson, K. A. Kiehl, Y. M. Wang, A. M. Michael, and V. D. Calhoun, “A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia,” *Magn Reson Mat Phys Biol Med*, vol. 23, no. 5, pp. 351–366, 2010.
- [227] L. K. Senel, T. Kilic, A. Gungor, E. Kopanoglu, H. E. Guven, E. U. Saritas, A. Koc, and T. Cukur, “Statistically segregated k-space sampling for accelerating multiple-acquisition mri,” *IEEE transactions on medical imaging*, vol. 38, no. 7, pp. 1701–1714, 2019.
- [228] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *Int Conf Mach Learn*, pp. 1310–1318, PMLR, 2013.
- [229] E. Ilicak, S. Cetin, E. Bulut, K. K. Oguz, E. U. Saritas, G. Unal, and T. Çukur, “Targeted vessel reconstruction in non-contrast-enhanced steady-state free precession angiography,” *NMR in Biomedicine*, vol. 29, no. 5, pp. 532–544, 2016.
- [230] T. Çukur, “Accelerated phase-cycled ssfp imaging with compressed sensing,” *IEEE transactions on medical imaging*, vol. 34, no. 1, pp. 107–115, 2014.
- [231] T. Çukur, M. Lustig, E. U. Saritas, and D. G. Nishimura, “Signal compensation and compressed sensing for magnetization-prepared mr angiography,” *IEEE transactions on medical imaging*, vol. 30, no. 5, pp. 1017–1027, 2011.
- [232] H. A. Bedel, I. Şivgin, and T. Çukur, “A graphical network layer for lagged analysis of fMRI data,” in *Sig Process Commun Appl Conf*, pp. 1–4, 2022.
- [233] O. Dalmaz, M. Yurt, and T. Çukur, “Resvit: residual vision transformers for multimodal medical image synthesis,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2598–2614, 2022.

- [234] Y. Gao and A. No, “Age estimation from fMRI data using recurrent neural network,” *Appl Sci*, vol. 12, no. 2, p. 749, 2022.
- [235] S. Arslan, S. I. Ktena, B. Glocker, and D. Rueckert, “Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity,” in *Graph Biomed Image Anal Integ Med Imaging Non-imaging Modal*, pp. 3–13, Springer, 2018.
- [236] T. Çukur, J. H. Lee, N. K. Bangerter, B. A. Hargreaves, and D. G. Nishimura, “Non-contrast-enhanced flow-independent peripheral mr angiography with balanced ssfp,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 61, no. 6, pp. 1533–1539, 2009.
- [237] D. Miwa, V. N. L. Duy, and I. Takeuchi, “Valid p-value for deep learning-driven salient region,” *arXiv:2301.02437*, 2023.
- [238] T. Yarkoni, R. A. Poldrack, T. E. Nichols, D. C. Van Essen, and T. D. Wager, “Large-scale automated synthesis of human functional neuroimaging data,” *Nat Methods*, vol. 8, no. 8, pp. 665–670, 2011.
- [239] T. Cukur, M. Lustig, and D. G. Nishimura, “Improving non-contrast-enhanced steady-state free precession angiography with compressed sensing,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 61, no. 5, pp. 1122–1131, 2009.
- [240] E. A. Allen, E. B. Erhardt, E. Damaraju, W. Gruner, J. M. Segall, R. F. Silva, M. Havlicek, S. Rachakonda, J. Fries, R. Kalyanam, *et al.*, “A baseline for the multivariate comparison of resting-state networks,” *Front Syst Neurosci*, vol. 5, p. 2, 2011.
- [241] R. A. Poldrack and T. Yarkoni, “From Brain Maps to Cognitive Ontologies: Informatics and the Search for Mental Structure,” *Ann. Rev. Psych.*, vol. 67, pp. 587–612, 2016.

- [242] M. S. Vendetti and S. A. Bunge, “Evolutionary and developmental changes in the lateral frontoparietal network: a little goes a long way for higher-level cognition.,” *Neuron*, vol. 84, no. 5, pp. 906–917, 2014.
- [243] M. D. Lieberman, M. A. Straccia, M. L. Meyer, M. Du, and K. M. Tan, “Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence,” *Neurosci. Biobehav. Rev.*, vol. 99, pp. 311–328, 2019.
- [244] F. Van Overwalle, Q. Ma, N. Haihambo, T. Bylemans, B. Catoira, M. Firouzi, M. Li, M. Pu, E. Heleven, C. Baeken, K. Baetens, and N. Deroost, “A Functional Atlas of the Cerebellum Based on NeuroSynth Task Coordinates,” *Cerebellum*, 2023.
- [245] T. Çukur, M. Yamada, W. R. Overall, P. Yang, and D. G. Nishimura, “Positive contrast with alternating repetition time ssfp (parts): A fast imaging technique for spio-labeled cells,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 63, no. 2, pp. 427–437, 2010.
- [246] R. Pope, S. Douglas, A. Chowdhery, J. Devlin, J. Bradbury, A. Levskaya, J. Heek, K. Xiao, S. Agrawal, and J. Dean, “Efficiently scaling transformer inference,” *arXiv:2211.05102*, 2022.
- [247] M. Li, W. Hsu, X. Xie, J. Cong, and W. Gao, “SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network,” vol. 39, no. 7, pp. 2289–2301, 2020.
- [248] T. Çukur and D. G. Nishimura, “Multiple repetition time balanced steady-state free precession imaging,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 62, no. 1, pp. 193–204, 2009.
- [249] M. J. Singleton, “Functional magnetic resonance imaging,” *The Yale Journal of Biology and Medicine*, vol. 82, no. 4, p. 233, 2009.

- [250] H. Jang, S. M. Plis, V. D. Calhoun, and J.-H. Lee, “Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks,” *NeuroImage*, vol. 145, pp. 314–328, 2017.
- [251] X. Wang, L. Yao, I. Rekik, and Y. Zhang, “Contrastive functional connectivity graph learning for population-based fmri classification,” in *MICCAI*, pp. 221–230, 2022.
- [252] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, “Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,” *Med Image Anal*, vol. 79, p. 102470, 2022.
- [253] T. Matsui, M. Taki, T. Q. Pham, J. Chikazoe, and K. Jimura, “Counterfactual explanation of brain activity classifiers using image-to-image transfer by generative adversarial network,” *Front Neuroinf*, vol. 15, p. 79, 2022.
- [254] P. Rodriguez, M. Caccia, A. Lacoste, L. Zamparo, I. Laradji, L. Charlin, and D. Vazquez, “Beyond trivial counterfactual explanations with diverse valuable explanations,” in *IEEE ICCV*, pp. 1056–1065, 2021.
- [255] L. Tomaz Da Silva, N. B. Esper, D. D. Ruiz, F. Meneguzzi, and A. Buchweitz, “Visual Explanation for Identification of the Brain Bases for Developmental Dyslexia on fMRI Data,” *Front Comput Neurosci*, vol. 15, p. 594659, 2021.
- [256] Y. Kazemi and S. Houghten, “A deep learning pipeline to classify different stages of Alzheimer’s disease from fMRI data,” in *IEEE CIBCB*, pp. 1–8, 2018.
- [257] B. J. Devereux, A. Clarke, and L. K. Tyler, “Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway,” *Sci Rep*, vol. 8, no. 1, p. 10636, 2018.

- [258] J. P. Cohen, R. Brooks, S. En, E. Zucker, A. Pareek, M. P. Lungren, and A. Chaudhari, “Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays,” *arXiv:2102.09475*, 2021.
- [259] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Adv Neural Inf Process Syst*, vol. 33, pp. 6840–6851, 2020.
- [260] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” *arXiv:2011.13456*, 2020.
- [261] P. Sanchez, A. Kascenas, X. Liu, A. Q. O’Neil, and S. A. Tsaftaris, “What is healthy? generative counterfactual diffusion for lesion localization,” in *DGM4MICCAI*, pp. 34–44, 2022.
- [262] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin, “Diffusion models for medical anomaly detection,” in *MICCAI*, vol. 13438, pp. 35–45, Springer, 2022.
- [263] W. H. L. Pinaya, M. S. Graham, R. Gray, P. F. Da Costa, P.-D. Tudosiu, P. Wright, Y. H. Mah, A. D. MacKinnon, J. T. Teo, R. Jager, D. Werring, G. Rees, P. Nachev, S. Ourselin, and M. J. Cardoso, “Fast Unsupervised Brain Anomaly Detection and Segmentation with Diffusion Models,” *arXiv:2206.03461*, 2022.
- [264] C. I. Bercea, M. Neumayr, D. Rueckert, and J. A. Schnabel, “Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models,” *arXiv preprint arXiv:2305.19643*, 2023.
- [265] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv:2010.02502*, 2020.
- [266] R. C. deCharms, “Applications of real-time fMRI,” *Nat Rev Neurosci*, vol. 9, no. 9, pp. 720–729, 2008.

- [267] S. Vieira, W. H. Pinaya, and A. Mechelli, “Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications,” *Neurosci Biobehav Rev*, vol. 74, pp. 58–75, 2017.
- [268] S. I. Ktena, S. Parisot, E. Ferrante, M. Rajchl, M. Lee, B. Glocker, and D. Rueckert, “Metric learning with spectral graph convolutions on brain connectivity networks,” *NeuroImage*, vol. 169, pp. 431–442, 2018.
- [269] I. Sivgin, H. A. Bedel, S. Öztürk, and T. Çukur, “A plug-in graph neural network to boost temporal sensitivity in fMRI analysis,” *IEEE J Biomed Health Inf*, 2024.
- [270] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, “Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results,” *Med Image Anal*, vol. 65, p. 101765, 2020.
- [271] C. A. Ellis, R. L. Miller, and V. D. Calhoun, “An Approach for Estimating Explanation Uncertainty in fMRI dFNC Classification,” in *IEEE BIBE*, pp. 297–300, 2022.
- [272] A. Gotsopoulos, H. Saarimäki, E. Glerean, I. P. Jääskeläinen, M. Sams, L. Nummenmaa, and J. Lampinen, “Reproducibility of importance extraction methods in neural network based fMRI classification,” *NeuroImage*, vol. 181, pp. 44–54, 2018.
- [273] H. Vu, H.-C. Kim, M. Jung, and J.-H. Lee, “fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations,” *NeuroImage*, vol. 223, p. 117328, 2020.
- [274] Q.-H. Lin, Y.-W. Niu, J. Sui, W.-D. Zhao, C. Zhuo, and V. D. Calhoun, “SSPNet: An interpretable 3D-CNN for classification of schizophrenia using phase maps of resting-state complex-valued fMRI data,” *Med Image Anal*, vol. 79, p. 102430, 2022.

- [275] Y. Li, C. Sun, P. Li, Y. Zhao, G. K. Mensah, Y. Xu, H. Guo, and J. Chen, “Hypernetwork construction and feature fusion analysis based on sparse group lasso method on fMRI dataset,” *Front Neurosci*, vol. 14, p. 60, 2020.
- [276] J. Hu, L. Cao, T. Li, B. Liao, S. Dong, and P. Li, “Interpretable learning approaches in resting-state functional connectivity analysis: the case of autism spectrum disorder,” *Comput Math Methods Med*, p. 1394830, 2020.
- [277] P. M. N. Dos Santos, S. L. Mendes, C. Biazoli, A. Gadelha, G. A. Salum, E. C. Miguel, L. A. Rohde, and J. R. Sato, “Assessing atypical brain functional connectivity development: An approach based on generative adversarial networks,” *Front Neurosci*, vol. 16, p. 1025492, 2023.
- [278] H. Reynaud, A. Vlontzos, M. Dombrowski, C. Gilligan Lee, A. Beqiri, P. Leeson, and B. Kainz, “D’artagnan: Counterfactual video generation,” in *MICCAI*, pp. 599–609, 2022.
- [279] N. Pawlowski, D. C. Castro, and B. Glocker, “Deep structural causal models for tractable counterfactual inference,” in *NeurIPS*, 2020.
- [280] M. Augustin, V. Boreiko, F. Croce, and M. Hein, “Diffusion visual counterfactual explanations,” in *NeurIPS*, 2022.
- [281] P. Sanchez and S. A. Tsaftaris, “Diffusion causal models for counterfactual estimation,” *arXiv:2202.10166*, 2022.
- [282] G. Jeanneret, L. Simon, and F. Jurie, “Diffusion models for counterfactual explanations,” in *ACCV*, pp. 858–876, 2022.
- [283] H. Chung, B. Sim, D. Ryu, and J. C. Ye, “Improving diffusion models for inverse problems using manifold constraints,” in *NeurIPS*, vol. 35, pp. 25683–25696, 2022.
- [284] T. Salimans and J. Ho, “Progressive distillation for fast sampling of diffusion models,” *arXiv:2202.00512*, 2022.
- [285] W. Luo, “A comprehensive survey on knowledge distillation of diffusion models,” *arXiv:2304.04262*, 2023.

- [286] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, *et al.*, “ediffi: Text-to-image diffusion models with an ensemble of expert denoisers,” *arXiv:2211.01324*, 2022.
- [287] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, “The minimal preprocessing pipelines for the human connectome project,” *NeuroImage*, vol. 80, pp. 105–124, 2013.
- [288] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen, “A multi-modal parcellation of human cerebral cortex,” *Nature*, vol. 536, no. 7615, pp. 171–178, 2016.
- [289] J. L. Lancaster, M. G. Woldorff, L. M. Parsons, M. Liotti, C. S. Freitas, L. Rainey, P. V. Kochunov, D. Nickerson, S. A. Mikiten, and P. T. Fox, “Automated talairach atlas labels for functional brain mapping,” *Hum Brain Map*, vol. 10, no. 3, pp. 120–131, 2000.
- [290] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, “Linformer: Self-Attention with Linear Complexity,” *arXiv:2006.04768*, 2020.
- [291] M. Atad, V. Dmytrenko, Y. Li, X. Zhang, M. Keicher, J. Kirschke, B. Wiestler, A. Khakzar, and N. Navab, “CheXplaining in Style: Counterfactual Explanations for Chest X-rays using StyleGAN,” *arXiv:2207.07553*, 2022.
- [292] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *ECCV*, pp. 818–833, Springer, 2014.
- [293] J. Castro, D. Gómez, and J. Tejada, “Polynomial calculation of the shapley value based on sampling,” *Comput Oper Res*, vol. 36, no. 5, pp. 1726–1730, 2009.
- [294] R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. T. Yeo, “The organization of the human cerebellum estimated by intrinsic functional connectivity,” *J Neurophys*, vol. 106, no. 5, pp. 2322–2345, 2011.

- [295] Y. Kamitani and F. Tong, “Decoding the visual and subjective contents of the human brain,” *Nat Neurosci*, vol. 8, no. 5, pp. 679–685, 2005.
- [296] T. Naselaris, K. Kay, S. Nishimoto, and G. J.L., “Encoding and decoding in fMRI,” *NeuroImage*, vol. 56, no. 2, pp. 400–410, 2011.
- [297] X. Zhang, M. Liang, W. Qin, B. Wan, C. Yu, and D. Ming, “Gender differences are encoded differently in the structure and function of the human brain revealed by multimodal mri,” *Front Hum Neurosci*, vol. 14, 2020.
- [298] E. McGlade, J. Rogowska, and D. Yurgelun-Todd, “Sex differences in orbitofrontal connectivity in male and female veterans with tbi,” *Brain Imaging Behav*, vol. 9, no. 3, pp. 535–549, 2015.
- [299] S. Weis, K. R. Patil, F. Hoffstaedter, A. Nostro, B. T. T. Yeo, and S. B. Eickhoff, “Sex Classification by Resting State Brain Connectivity,” *Cereb Cortex*, vol. 30, no. 2, pp. 824–835, 2019.
- [300] D. Tomasi and N. Volkow, “Laterality patterns of brain functional connectivity: gender effects,” *Cereb Cortex*, vol. 22, no. 6, pp. 1455–1462, 2012.
- [301] B. Du, X. Cheng, Y. Duan, and H. Ning, “fMRI Brain Decoding and Its Applications in Brain-Computer Interface: A Survey,” *Brain Sci*, vol. 12, no. 2, 2022.
- [302] I. Kiremitci, O. Yilmaz, E. Celik, M. Shahdloo, A. G. Huth, and T. Çukur, “Attentional Modulation of Hierarchical Speech Representations in a Multitalker Environment,” *Cereb Cortex*, vol. 31, no. 11, pp. 4986–5005, 2021.
- [303] D. Wang, F. Fan, Z. Wu, R. Liu, F. Wang, and H. Yu, “CTformer: convolution-free Token2Token dilated vision transformer for low-dose CT denoising,” *Phys Med Biol*, vol. 68, no. 6, p. 065012, 2023.
- [304] T. Çukur, M. Lustig, and D. G. Nishimura, “Multiple-profile homogeneous image combination: application to phase-cycled ssfp and multicoil imaging,” *Magnetic resonance in medicine*, vol. 60, no. 3, pp. 732–738, 2008.
- [305] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “DiffWave: A versatile diffusion model for audio synthesis,” *arXiv:2009.09761*, 2021.

- [306] Z.-X. Cui, C. Liu, C. Cao, Y. Liu, J. Cheng, Q. Zhu, Y. Zhu, H. Wang, and D. Liang, “Meta-learning enabled score-based generative model for 1.5t-like image reconstruction from 0.5t mri,” *arXiv:2305.02509*, 2023.
- [307] T. Çukur, N. K. Bangerter, and D. G. Nishimura, “Enhanced spectral shaping in steady-state free precession imaging,” *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 58, no. 6, pp. 1216–1223, 2007.
- [308] W. Xia, Q. Lyu, and G. Wang, “Low-dose CT using denoising diffusion probabilistic model for $20\times$ speedup,” *arXiv:2209.15136*, 2022.
- [309] Y. Fan, H. Rao, H. Hurt, J. Giannetta, M. Korczykowski, D. Shera, B. B. Avants, J. C. Gee, J. Wang, and D. Shen, “Multivariate examination of brain abnormality using both structural and functional MRI,” *NeuroImage*, vol. 36, no. 4, pp. 1189–1199, 2007.
- [310] L. Zhang, L. Wang, J. Gao, S. L. Risacher, J. Yan, G. Li, T. Liu, and D. Zhu, “Deep fusion of brain structure-function in mild cognitive impairment,” *Med Image Anal*, vol. 72, p. 102082, 2021.
- [311] A. Craik, Y. He, and J. L. Contreras-Vidal, “Deep learning for electroencephalogram (EEG) classification tasks: a review,” *J Neural Eng*, vol. 16, no. 3, p. 031001, 2019.
- [312] J. Aoe, R. Fukuma, T. Yanagisawa, T. Harada, M. Tanaka, M. Kobayashi, Y. Inoue, S. Yamamoto, Y. Ohnishi, and H. Kishima, “Automatic diagnosis of neurological diseases using MEG signals with a deep neural network,” *Scientific Reports*, vol. 9, no. 1, p. 5057, 2019.
- [313] C. Eastmond, A. Subedi, S. De, and X. Intes, “Deep learning in fNIRS: a review,” *Neurophotonics*, vol. 9, no. 4, p. 041411, 2022.
- [314] A. P. Buccino, M. Kordovan, T. V. Ness, B. Merkt, P. D. Häfliger, M. Fyhn, G. Cauwenberghs, S. Rotter, and G. T. Einevoll, “Combining biophysical modeling and deep learning for multielectrode array neuron localization and classification,” *J Neurophysiol*, vol. 120, no. 3, pp. 1212–1232, 2018.

- [315] W. Yan, G. Qu, W. Hu, A. Abrol, B. Cai, C. Qiao, S. M. Plis, Y.-P. Wang, J. Sui, and V. D. Calhoun, “Deep Learning in Neuroimaging: Promises and challenges,” *IEEE Sig Process Mag*, vol. 39, no. 2, pp. 87–98, 2022.

