


A Plug-In Graph Neural Network to Boost Temporal Sensitivity in fMRI Analysis

Irmak Sivgin, Hasan Atakan Bedel, Saban Ozturk, and Tolga Çukur , Senior Member, IEEE

Abstract—Learning-based methods offer performance leaps over traditional methods in classification analysis of high-dimensional functional MRI (fMRI) data. In this domain, deep-learning models that analyze functional connectivity (FC) features among brain regions have been particularly promising. However, many existing models receive as input temporally static FC features that summarize inter-regional interactions across an entire scan, reducing the temporal sensitivity of classifiers by limiting their ability to leverage information on dynamic FC features of brain activity. To improve the performance of baseline classification models without compromising efficiency, here we propose a novel plug-in based on a graph neural network, GraphCorr, to provide enhanced input features to baseline models. The proposed plug-in computes a set of latent FC features with enhanced temporal information while maintaining comparable dimensionality to static features. Taking brain regions as nodes and blood-oxygen-level-dependent (BOLD) signals as node inputs, GraphCorr leverages a node embedder module based on a transformer encoder to capture dynamic latent representations of BOLD signals. GraphCorr also leverages a lag filter module to account for delayed interactions across nodes by learning correlational features of windowed BOLD signals across time delays. These two feature groups are then fused via a message

passing algorithm executed on the formulated graph. Comprehensive demonstrations on three public datasets indicate improved classification performance for several state-of-the-art graph and convolutional baseline models when they are augmented with GraphCorr.

Index Terms—Connectivity, functional MRI, graph, neural network, time series.

I. INTRODUCTION

THE human brain comprises functional networks composed of multiple brain regions that interactively process information to mediate cognitive processes [1]. In turn, correlated activity among brain regions within individual functional networks has been associated with unique mental states [2], [3], [4], [5]. 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 [4], [6], [7]. In fMRI studies, networks are commonly assessed via functional connectivity (FC) measures that reflect similarity of BOLD signals among brain regions [7], [8], [9]. 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) [10], [11], [12], [13]. Unfortunately, conventional methods often yield poor capture of intricate information patterns in whole-brain fMRI time series [14].

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 [15], [16], [17], [18], [19]. Earlier attempts in this domain have proposed shallow multi-layer perceptron (MLP) [20], [21] and Boltzmann machine (BM) models [15], [22]. Later studies have adopted deeper architectures based on convolutional neural network (CNN) [16], [23], [24], graph neural network (GNN) [8], [17], [25], [26], [27], [28], [29], [30], and transformer [14], [31], [32], [33], [34] models for improved performance. These models construct a set of nodes corresponding to brain regions defined based on an atlas [29], [35], [36], [37], and receive input features at these nodes based on the FC strength among brain regions [17], [38]. A pervasive approach has been to employ static FC features derived from summary correlation measures across the entire fMRI scan [17], [39]. Yet, this approach is often insufficiently sensitive to the dynamic inter-regional interactions during resting-state or cognitive tasks [40]. While alternative strategies have recently been

Manuscript received 20 June 2023; revised 22 February 2024 and 29 April 2024; accepted 12 June 2024. Date of publication 17 June 2024; date of current version 6 September 2024. The work of Tolga Çukur was supported in part by TUBITAK BİDEB scholarship awarded to Hasan Atakan Bedel, in part by TUBITAK under Grant 121N029 in part by the European Joint Programme Neurodegenerative Disease Research (JPND) 2020 call, and in part by novel imaging and brain stimulation methods and technologies related to Neurodegenerative Diseases for the Neuripides Project Neurofeedback for self-stimulation of the brain as therapy for Parkinson Disease. The Neuripides Project was supported in part by JPND: The Netherlands, The Netherlands Organization for Health Research and Development (ZonMw), in part by Germany, Federal Ministry of Education and Research (BMBF), in part by Czech Republic, Ministry of Education, Youth and Sports (MEYS), in part by France, French National Research Agency (ANR), in part by Canada, Canadian Institutes of Health Research (CIHR), and in part by Turkey, Scientific and Technological Research Council of Turkey (TUBITAK). (Irmak Sivgin and Hasan Atakan Bedel contributed equally to this work.) (Corresponding author: Tolga Çukur.)

Irmak Sivgin, Hasan Atakan Bedel, and Tolga Çukur are with the Department of Electrical-Electronics Engineering, National Magnetic Resonance Research Center (UMRAM), Bilkent University, 06800 Ankara, Türkiye (e-mail: sivginirmak@gmail.com; abedel@ee.bilkent.edu.tr; cukur@ee.bilkent.edu.tr).

Saban Ozturk is with the Department of Electrical-Electronics Engineering, National Magnetic Resonance Research Center (UMRAM), Bilkent University, 06800 Ankara, Türkiye, and also with Ankara Haci Bayram Veli University, 06570 Ankara, Türkiye (e-mail: sabn.ozturk@gmail.com).

Digital Object Identifier 10.1109/JBHI.2024.3415000

proposed to capture the temporal variability in FC features, these methods commonly consider instantaneous correlations¹ across local time windows within the time series [41], [42], [43]. As such, they lack explicit mechanisms to capture delayed correlations² eminent in fMRI data due to hierarchical cognitive processing or hemodynamic lags in BOLD signals [44].

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 [38], [45], [46]. Thus, we first examine gender detection on resting-state scans in the HCP-Rest dataset from Human Connectome Project [47] and on movie-watching fMRI scans in the ID1000 dataset from Amsterdam Open MRI Collection [48]. Functional connectivity differences are not exclusive to resting state, but they have also been reported during intentful execution of various cognitive tasks [30], [41], [49]. Thus, we also examine cognitive-task detection on task-based fMRI scans in the HCP-Task dataset [47].

GraphCorr is used as a plug-in to augment state-of-the-art learning-based fMRI classifiers including graphical (SAGE [50], GCN [27]), and convolutional (BrainNetCNN [16]) 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.

II. RELATED WORK

Cognitive processes in the human brain elicit broadly distributed response patterns spanning across multiple brain regions [51], which can be analyzed to classify stimulus or task

variables [52], [53]. Many analysis methods for fMRI rely on feature selection procedures to cope with the intrinsically high dimensionality of fMRI data [40], [54], [55]. Arguably, FC measures among brain regions have been the most commonly used feature set [16], [17], [20], [38]. Previous studies have reported that external variables or disease states can be detected given FC features of individual subjects under resting state [56], [57], [58], cognitive tasks [10], [17], or both [13]. Given their success in fMRI analysis, FC features have also been adopted in recent DL methods [59], [60]. 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 [15], [20], [21], [22], or deep convolutional architectures [16], [24]. Later studies have considered GNN models given their natural fit to analyzing fMRI data that follows an intrinsic graph-like connectivity structure [17], [30], [61], [62], [63], [64]. 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 [11], [24].

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 [8], [42], [43], [65], [66]. 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 [32], [41]. 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 [41]. Besides elevated model complexity, these methods can suffer from limitations of RNNs in terms of vanishing/exploding gradients [67], [68]. Importantly, these previous methods primarily focus on temporal variations in instantaneous correlations, while neglecting delayed inter-regional correlations [44].

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 [16], [17], 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 [8], [41], [69], 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

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

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

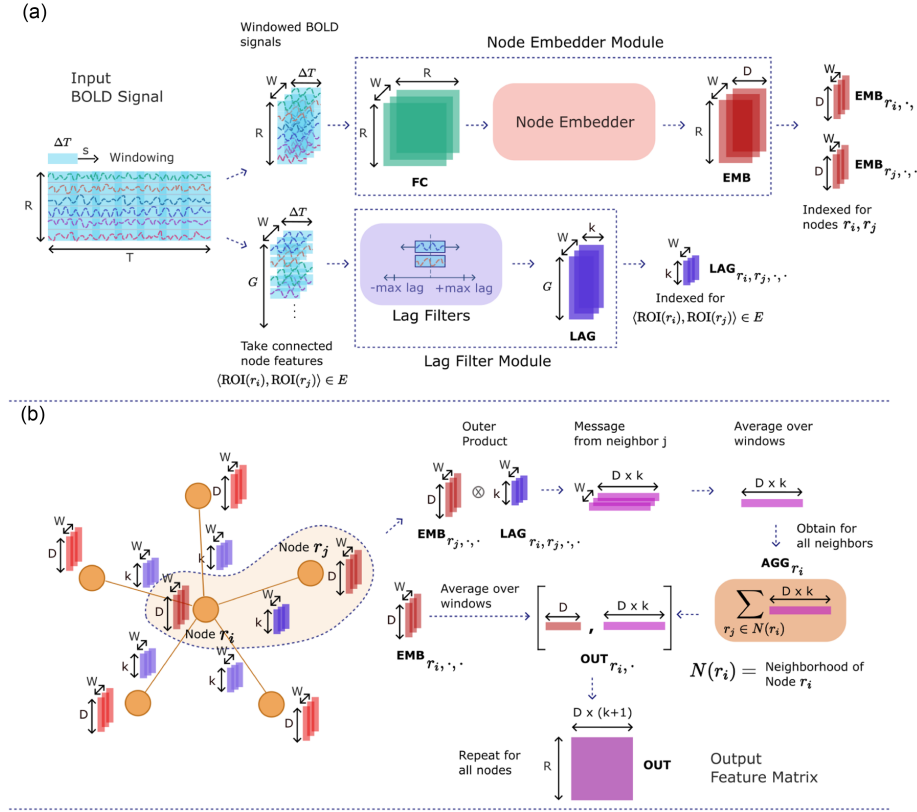


Fig. 1. Overview of GraphCorr. (a) GraphCorr utilizes two modules in parallel to extract dynamic, lagged features of inter-regional correlations across the brain. The node embedder module receives as input time-windowed BOLD signals, and uses a transformer encoder to compute node embeddings of dynamic 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 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, $\mathbf{OUT} \in \mathbb{R}^{R \times (D+k)}$.

architectures [41], it leverages a transformer encoder on dynamic FC features that enables efficient parallel processing. Unlike methods that solely focus on instantaneous correlations [16], 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.

III. 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 [17], [41]. 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 models.³ 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. 1). The methodological components and procedures are described below. Code for GraphCorr will be shared at <https://github.com/icon-lab/GraphCorr> upon publication.

A. 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 $N = \{\text{ROI}(r_i) \mid r_i = 1, \dots, R\}$ includes a total of R ROIs defined according to the atlas. Meanwhile, the binary edge set is taken as $E = \{(\text{ROI}(r_i), \text{ROI}(r_j)) \mid r_i = 1, \dots, R; r_j \in \mathcal{N}(r_i)\}$ 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 [17]. An initial tensor of node features

³see [70] for a preliminary version of this work presented at SIU 2022

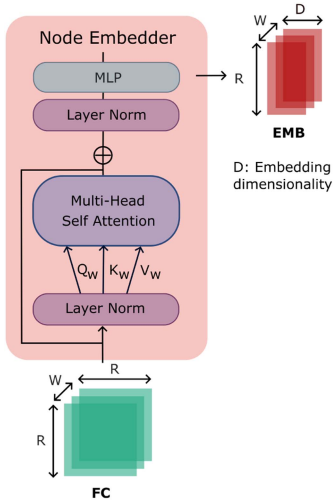


Fig. 2. The node embedder module. The module input is a tensor of time-windowed dynamic FC features $\mathbf{dFC} \in \mathbb{R}^{R \times R \times W}$, with R denoting the number of ROIs, W denoting the number of time windows. The input tensor is processed via a transformer encoder with multi-head self-attention (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.

$\mathbf{F} = \{f_{r_i} \mid r_i = 1, \dots, R\}$ 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 :

$$W = \left\lfloor \frac{T - \Delta T}{s} \right\rfloor. \quad (1)$$

For a given node 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 $\mathbf{F} \in \mathbb{R}^{R \times \Delta T \times W}$.

B. 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. 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 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 $U_q, U_k, U_v \in \mathbb{R}^{R \times d}$:

$$Q_w = \text{LN}([\mathbf{dFC}_{r_1, \cdot, w}, \mathbf{dFC}_{r_2, \cdot, w}, \dots, \mathbf{dFC}_{r_R, \cdot, w}]) U_q,$$

$$K_w = \text{LN}([\mathbf{dFC}_{r_1, \cdot, w}, \mathbf{dFC}_{r_2, \cdot, w}, \dots, \mathbf{dFC}_{r_R, \cdot, w}]) U_k,$$

$$V_w = \text{LN}([\mathbf{dFC}_{r_1, \cdot, w}, \mathbf{dFC}_{r_2, \cdot, w}, \dots, \mathbf{dFC}_{r_R, \cdot, w}]) U_v, \quad (2)$$

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 [71]:

$$\mathbf{A}_w = \text{Att}(Q_w, K_w, V_w) = \text{Softmax} \left(\frac{Q_w K_w^\top}{\sqrt{d}} \right) V_w, \quad (3)$$

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:

$$\mathbf{EMB}_{\cdot, \cdot, w} = \text{MLP}(\text{LN}(\mathbf{A}_w)) = \text{GELU}(\text{LN}(\mathbf{A}_w) \mathbf{M}_1) \mathbf{M}_2, \quad (4)$$

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 (4), $\mathbf{EMB} \in \mathbb{R}^{R \times D \times W}$ denotes the output embedding tensor with D taken as embedding dimensionality such that $D < R$.

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. 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:

$$\mathbf{X} = [\mathbf{0}_{(R \times m \times W)}, \mathbf{F}_{(R \times \Delta T \times W)}, \mathbf{0}_{(R \times m \times W)}], \quad (5)$$

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:

$$\rho_{r_i, r_j, w, \tau} = (\mathbf{X}_{r_i, \cdot, w} \star \mathbf{X}_{r_j, \cdot, w})[\tau], \quad (6)$$

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:

$$\mathbf{LAG}_{r_i, r_j, w, \cdot} = \text{GELU}(\rho_{r_i, r_j, w, \cdot} \mathbf{M}_{LF}) \quad (7)$$

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

C. Graph Learning

The node embedder module produces an embedding tensor, \mathbf{EMB} , that reflects instantaneous inter-regional correlations between graph nodes, whereas the lag filter produces a lag activation tensor, \mathbf{LAG} , that reflects delayed inter-regional correlations between nodes. To consolidate these two feature sets

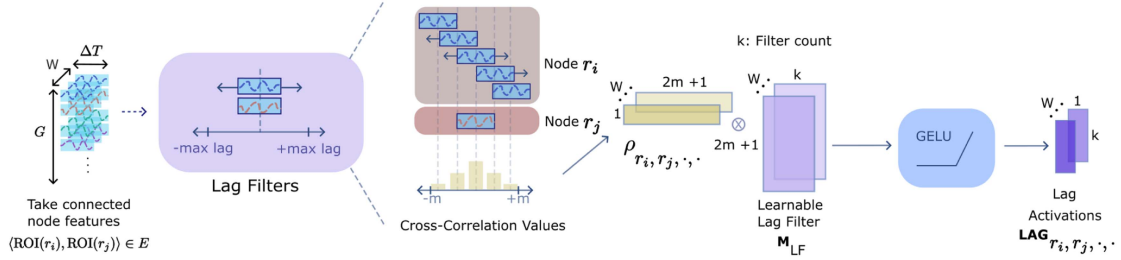


Fig. 3. The lag filter module. The module input is a tensor of time-windowed BOLD signals $\mathbf{F} \in \mathbb{R}^{R \times \Delta T \times W}$, with ΔT denoting the duration of each window. Delayed connectivity features of pairs of ROIs are computed as the cross-correlation of their BOLD signals, 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 $\mathbf{LAG} \in \mathbb{R}^{R \times R \times W \times k}$.

on the graph, node embeddings are taken as node features and lag activations are taken as edge weights (Fig. 1). To compute enhanced features, a message passing algorithm is then run on the graph to fuse the feature sets [61]. 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:

$$\mathbf{MES}_{r_i, r_j, \cdot, w} = \text{flatten}(\mathbf{EMB}_{r_i, \cdot, w} \otimes \mathbf{LAG}_{r_i, r_j, w, \cdot}), \quad (8)$$

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:

$$\mathbf{AGG}_{r_i, \cdot} = \sum_{r_j \in \mathcal{N}(r_i)} \frac{1}{W} \sum_{w=1}^W \mathbf{MES}_{r_i, r_j, \cdot, w}, \quad (9)$$

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 ROI(r_i):

$$\mathbf{OUT}_{r_i, \cdot} = \left[\frac{1}{W} \sum_{w=1}^W \mathbf{EMB}_{r_i, \cdot, w}, \mathbf{AGG}_{r_i, \cdot} \right]. \quad (10)$$

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.

IV. EXPERIMENTS

A. Experimental Procedures

Demonstrations were performed on fMRI data from the HCP S1200 release⁴ [47] and ID1000 dataset from Amsterdam Open MRI Collection (AOMIC)⁵ [48]. 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 [72], and the AAL atlas that provides anatomy-based definitions of $R = 116$ ROIs [35]. 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, hyperparameters were selected to maximize the average performance across the validation sets [17]. A common set of hyperparameters that were observed to yield near-optimal performance were used across datasets and atlases.

B. 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 [17]. Hyperparameters of both baseline models and plug-in methods were selected to optimize validation performance. Implementation details are given below.

1) Baseline Models: Several graphical and convolutional baseline models were examined for gender detection and cognitive-task detection from fMRI scans.

⁴[Online]. Available: <https://db.humanconnectome.org>

⁵[Online]. Available: <https://openneuro.org/datasets/ds003097/versions/1.2.1>

SAGE: SAGE is a GNN model based on modules containing graph convolution, pooling and fully-connected layers [50]. 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 3×10^{-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 [27]. 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 5×10^{-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 [16]. A hidden dimension of 32 and a dropout rate of 0.1 were used. Cross-validated hyperparameters were a learning rate of 2×10^{-4} , 20 epochs, and a batch size of 16.

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 [69]. 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 [73]. 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.

C. 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 [38], [74].

For the vanilla variant, gradients were computed with respect to the input static FC features \mathbf{sFC} :

$$\mathbf{SAL}_{r_i, r_j}^{van} = |\nabla_{\mathbf{sFC}_{r_i, r_j}} y|, \quad (11)$$

where $\mathbf{SAL}^{van} \in \mathbb{R}^{R \times R}$ in 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:

$$\mathbf{rSAL}^{van} = \sum_{r_j=1}^R \mathbf{SAL}_{\cdot, r_j}^{van}, \quad (12)$$

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

$$\mathbf{SAL}_{r_i, r_j, w}^{aug} = |\nabla_{\mathbf{dFC}_{r_i, r_j, w}} y|, \quad (13)$$

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:

$$\mathbf{rSAL}^{aug} = \sum_{r_j=1}^R \left(\frac{1}{W} \sum_{w=1}^W \mathbf{SAL}_{\cdot, r_j, w}^{aug} \right). \quad (14)$$

For reliable inference, significant ROIs with absolute gradient values greater than zero were determined via a Wilcoxon signed-rank test across the test set [75]. 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 [76]. 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%.

V. RESULTS

A. Comparison Studies

GraphCorr was demonstrated on several state-of-the-art baseline classification models for fMRI analysis including SAGE [50], GCN [27], and BrainNetCNN [16]. 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 I for the Schaefer atlas, and in Table II 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 ($p < 0.05$),

TABLE I

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 GRAPH CORR AUGMENTED (GRAPH CORR) VARIANTS

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\pm0.0
	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\pm0.7	94.3\pm2.0	81.7\pm2.7	87.0\pm2.1	99.2\pm0.2	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\pm0.0
	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\pm0.1
	GraphCorr	89.9\pm2.2	94.5\pm1.6	80.8\pm1.0	87.9\pm1.8	99.3\pm0.2	99.8\pm0.1
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\pm2.6	94.7\pm1.9	82.7\pm1.6	89.9\pm2.2	98.8\pm0.2	99.9\pm0.0

Boldface indicates the top-performing plug-in for each model.

TABLE II

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 \pm 3.3	75.7 \pm 1.3	62.8 \pm 1.8	67.2 \pm 2.7	81.0 \pm 1.1	97.2 \pm 0.1
	Dyn. FC	69.0 \pm 3.2	76.4 \pm 2.3	67.1 \pm 4.0	72.6 \pm 2.9	87.3 \pm 0.9	98.6 \pm 0.2
	GRU	76.2 \pm 2.3	82.7 \pm 1.4	68.6 \pm 3.3	73.7 \pm 2.0	95.3 \pm 1.4	99.6\pm0.2
	GraphCorr	85.5\pm3.6	91.2\pm2.6	77.5\pm3.7	84.9\pm1.9	95.8\pm0.4	99.6\pm0.1
GCN	Vanilla	69.9 \pm 1.4	75.9 \pm 1.1	65.5 \pm 1.4	71.0 \pm 1.2	77.5 \pm 1.0	95.9 \pm 0.2
	Dyn. FC	69.4 \pm 2.3	75.9 \pm 2.6	68.4 \pm 4.6	74.8 \pm 3.3	85.0 \pm 1.5	98.5 \pm 0.1
	GRU	74.5 \pm 2.9	82.8 \pm 1.9	67.2 \pm 2.2	72.0 \pm 2.7	94.9 \pm 0.6	99.7 \pm 0.0
	GraphCorr	84.4\pm3.2	89.0\pm2.8	79.4\pm3.5	85.4\pm2.5	96.1\pm0.4	99.9\pm0.1
BrainNetCNN	Vanilla	68.2 \pm 3.5	74.8 \pm 2.1	75.0 \pm 2.2	81.4 \pm 3.3	84.7 \pm 0.1	97.8 \pm 0.2
	Dyn. FC	76.7 \pm 1.7	83.6 \pm 1.0	76.3 \pm 3.1	83.3 \pm 2.4	91.3 \pm 0.5	99.1 \pm 0.1
	GRU	73.4 \pm 2.2	81.8 \pm 1.9	75.1 \pm 2.2	82.9 \pm 2.1	67.5 \pm 2.8	95.4 \pm 0.6
	GraphCorr	84.0\pm2.9	91.3\pm2.7	79.0\pm1.8	86.2\pm0.8	95.4\pm0.6	99.7\pm0.1

Boldface indicates the top-performing plug-in for each model.

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.

B. 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 I–II). 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 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 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 [38], [46]. 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

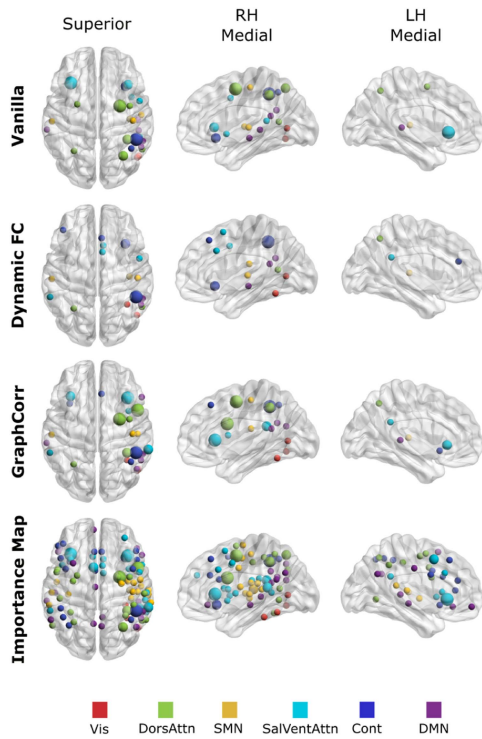


Fig. 4. Explanatory analysis for gender detection classifiers. Saliency maps are shown for vanilla, dynamic FC- and GraphCorr-augmented SAGE models on the HCP-Rest dataset, along with reference importance maps derived via a meta-analysis of neuroimaging literature. Important ROIs are marked with colored 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.).

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 [45], [46]. Previous studies have reported gender differences in these cognitive abilities, and in activations across associated regions [46], [77].

Fig. 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 [46]. 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

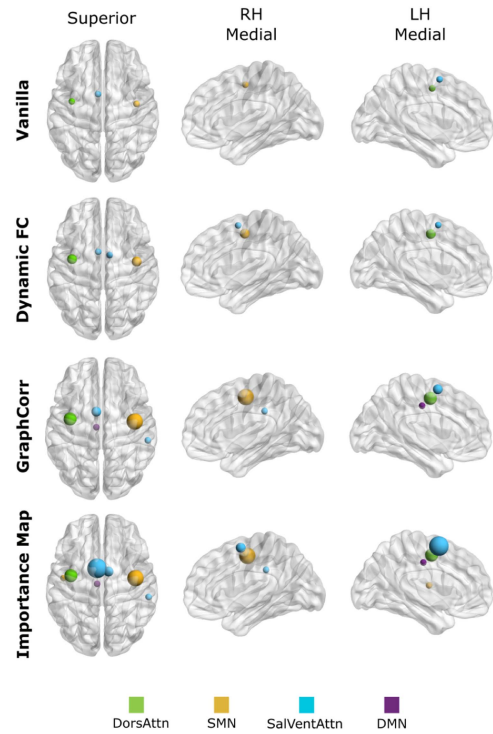


Fig. 5. Explanatory analysis for cognitive-task detection classifiers. Saliency maps for the motor task are shown based on vanilla, dynamic FC- and GraphCorr-augmented SAGE models on the HCP-Task dataset, along with reference importance maps derived via a meta-analysis of neuroimaging literature.

network (cingulate, supramarginal) are assumed to be involved in motor control and association of motor and sensory data [38]. Taken together, these results suggest that GraphCorr elicits explanations that are more closely aligned with neuroscience findings in the literature.

C. 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 III 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.

TABLE III

PERFORMANCE FOR GRAPH CORR 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 Embedder	Lag Filter	Windowing	Acc (%)	ROC (%)
✗	✗	✗	75.2±2.8	85.3±1.7
✓	✗	✗	80.5±1.9	88.5±1.5
✓	✓	✗	81.2±1.7	88.7±2.1
✓	✓	✓	89.6±0.7	94.3±2.0

Boldface indicates the top-performing variant.

TABLE IV

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

Boldface indicates the top-performing variant.

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 IV 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 $s = 50$.

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 [47]. To obtain a basis for time-varying FC

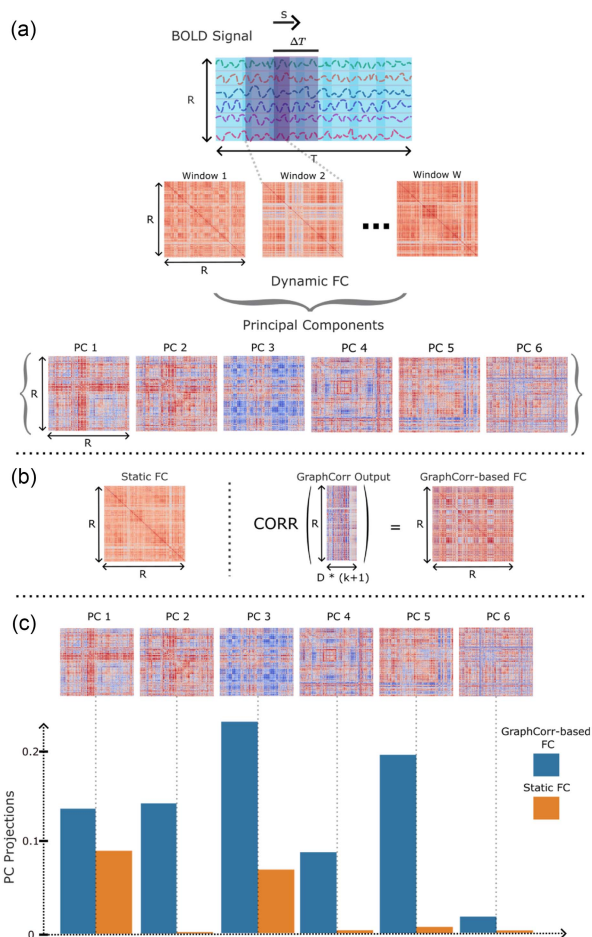


Fig. 6. Analysis of temporal sensitivity for GraphCorr on HCP-Task with the Schaefer atlas. (a) As reference for time-varying FC patterns in fMRI scans, conventional dynamic FC (dFC) features were computed via time-windowed correlation measures on BOLD signals. Principal components analysis (PCA) was 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.

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. 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.

VI. 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 [78], so it serves as a reliable benchmark to interpret neuroimaging findings [79], [80], [81]. 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 [40], [53]. 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 [16], [17] or continuous variables related to stimulus or task features [1]. An alternative procedure to examine cortical function rests on encoding models that instead predict BOLD signals from experimental variables [44], [49], [82]. 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 [83]. Learning might also be enhanced via data augmentation procedures that can produce a large variety of realistic samples from a learned distribution [84], [85].

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 [33]. 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 [86].

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 [87]. 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 [53].

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 [60], [88], [89].

VII. CONCLUSION

In this study, we introduced a novel plug-in graph neural network, GraphCorr, to improve the performance of learning-based models for fMRI classification. GraphCorr employs node embedder and lag filter modules to sensitively extract dynamic and lagged functional connectivity features from whole-brain fMRI time series. As such, it transforms raw BOLD signals into an efficient graph representation where neighboring nodes are taken as brain regions with correlated signals and node features are extracted via message passing on connectivity features from the two modules. This procedure restores the fine-grained temporal information that can otherwise be diminished in conventional functional connectivity features. As augmenting baseline classification models with GraphCorr significantly improves their performance and interpretability, GraphCorr holds great promise for analysis of fMRI time series.

REFERENCES

- [1] T. Çukur et al., "Attention during natural vision warps semantic representation across the human brain," *Nat. Neurosci.*, vol. 16, no. 6, pp. 763–770, 2013.
- [2] G. Lanza et al., "Preserved transcallosal inhibition to transcranial magnetic stimulation in nondemented elderly patients with leukoaraiosis," *BioMed. Res. Int.*, vol. 2013, 2013, Art. no. 351680.
- [3] R. Bella et al., "Enhanced motor cortex facilitation in patients with vascular cognitive impairment-no dementia," *Neurosci. Lett.*, vol. 503, no. 3, pp. 171–175, 2011.
- [4] D. Zhu et al., "Fusing DTI and fMRI data: A survey of methods and applications," *NeuroImage*, vol. 102, pp. 184–191, 2014.
- [5] 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.
- [6] 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.
- [7] K. Li et al., "Review of methods for functional brain connectivity detection using fMRI," *Comput. Med. Imag. Graph.*, vol. 33, no. 2, pp. 131–139, 2009.
- [8] S. Gadgil et al., "Spatio-temporal graph convolution for resting-state fMRI analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2020, pp. 528–538.
- [9] J. Pan et al., "Characterization multimodal connectivity of brain network by hypergraph GAN for alzheimer's disease analysis," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis.*, 2021, pp. 467–478.
- [10] J. Mourao-Miranda et al., "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.
- [11] B. Rashid et al., "Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity," *NeuroImage*, vol. 134, pp. 645–657, 2016.
- [12] N. U. Dosenbach et al., "Prediction of individual brain maturity using fMRI," *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010.
- [13] M. D. Rosenberg et al., "A neuromarker of sustained attention from whole-brain functional connectivity," *Nat. Neurosci.*, vol. 19, no. 1, pp. 165–171, 2016.
- [14] H. A. Bedel et al., "BoIT: Fused window transformers for fMRI time series analysis," *Med. Image Anal.*, vol. 88, 2023, Art. no. 102841.
- [15] S. M. Plis et al., "Deep learning for neuroimaging: A validation study," *Front Neurosci.*, vol. 8, 2014, Art. no. 229.
- [16] J. Kawahara et al., "BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment," *NeuroImage*, vol. 146, pp. 1038–1049, 2017.
- [17] X. Li et al., "BrainGNN: Interpretable brain graph neural network for fMRI analysis," *Med. Image Anal.*, vol. 74, 2021, Art. no. 102233.
- [18] A. S. Heinsfeld et al., "Identification of autism spectrum disorder using deep learning and the ABIDE dataset," *NeuroImage: Clin.*, vol. 17, pp. 16–23, 2018.
- [19] J. Kim et al., "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.
- [20] H. Shen et al., "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.
- [21] 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. 10th ACM Int. Conf. Bioinf. Comput. Biol. Health Inform.*, 2019, pp. 646–651.
- [22] R. D. Hjelm et al., "Restricted Boltzmann machines for neuroimaging: An application in identifying intrinsic networks," *NeuroImage*, vol. 96, pp. 245–260, 2014.
- [23] Y. Zhao et al., "Automatic recognition of fMRI-derived functional networks using 3-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1975–1984, Sep. 2018.
- [24] R. J. Meszlényi, K. Buza, and Z. Vidnyánszky, "Resting state fMRI functional connectivity-based classification using a convolutional neural network architecture," *Front. Neuroinform.*, vol. 11, 2017, Art. no. 61.
- [25] L. Shao et al., "Classification of ASD based on fMRI data with deep learning," *Cogn. Neurodynamics*, vol. 15, no. 6, pp. 961–974, 2021.
- [26] M. Saeidi et al., "Decoding task-based fMRI data with graph neural networks, considering individual differences," *Brain Sci.*, vol. 12, no. 8, 2022, Art. no. 1094.
- [27] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. Int. Conf. Learn. Rep.*, 2017.
- [28] G. Qu et al., "Brain functional connectivity analysis via graphical deep learning," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 5, pp. 1696–1706, May 2022.
- [29] 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, May 2023.
- [30] X. Li et al., "Graph neural network for interpreting task-fMRI biomarkers," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2019, pp. 485–493.
- [31] X. Yu et al., "Disentangling spatial-temporal functional brain networks via twin-transformers," 2022, *arXiv:2204.09225*.
- [32] I. Malkiel et al., "Pre-training and fine-tuning transformers for fMRI prediction tasks," 2021, *arXiv:2112.05761*.
- [33] S. Nguyen et al., "Attend and decode: 4D fMRI task state decoding using attention models," in *Proc. Mach. Learn. Health*, 2020, pp. 267–279.
- [34] W. Dai et al., "Brainformer: A hybrid CNN-transformer model for brain fMRI data classification," 2022, *arXiv:2208.03028*.
- [35] N. Tzourio-Mazoyer et al., "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.
- [36] 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.

- [37] A. R. Laird et al., "ALE meta-analysis workflows via the brainmap database: Progress towards a probabilistic functional brain atlas," *Front. Neuroinf.*, vol. 3, 2009, Art. no. 23.
- [38] B.-H. Kim and J. C. Ye, "Understanding graph isomorphism network for rs-fMRI functional connectivity analysis," *Front. Neurosci.*, vol. 14, 2020, Art. no. 545464.
- [39] J. Gan et al., "Brain functional connectivity analysis based on multi-graph fusion," *Med. Image Anal.*, vol. 71, 2021, Art. no. 102057.
- [40] K. A. Norman et al., "Beyond mind-reading: Multi-voxel pattern analysis of fMRI data," *Trend Cogn. Sci.*, vol. 10, no. 9, pp. 424–430, 2006.
- [41] B.-H. Kim, J. C. Ye, and J.-J. Kim, "Learning dynamic graph representation of brain connectome with spatio-temporal attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 4314–4327.
- [42] 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 Behav.*, vol. 9, no. 4, 2019, Art. no. e01255.
- [43] D. A. Handwerker et al., "Periodic changes in fMRI connectivity," *NeuroImage*, vol. 63, no. 3, pp. 1712–1719, 2012.
- [44] E. Celik et al., "Cortical networks of dynamic scene category representation in the human brain," *Cortex*, vol. 143, pp. 127–147, 2021.
- [45] S. J. Ritchie et al., "Sex differences in the adult human brain: Evidence from 5216 U.K. biobank participants," *Cereb. Cortex*, vol. 28, no. 8, pp. 2959–2975, 2018.
- [46] C. Zhang et al., "Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity," *Hum. Brain Mapping*, vol. 39, no. 4, pp. 1765–1776, 2018.
- [47] D. C. Van Essen et al., "The WU-minn human connectome project: An overview," *NeuroImage*, vol. 80, pp. 62–79, 2013.
- [48] L. Snoek et al., "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.
- [49] 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.
- [50] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1025–1035.
- [51] M. W. Woolrich et al., "Temporal autocorrelation in univariate linear modeling of fMRI data," *NeuroImage*, vol. 14, no. 6, pp. 1370–1386, 2001.
- [52] D. Z. Bolling et al., "Enhanced neural responses to rule violation in children with autism: A comparison to social exclusion," *Devlop. Cogn. Neurosci.*, vol. 1, no. 3, pp. 280–294, 2011.
- [53] M. Shahdloo, E. Çelik, and T. Çukur, "Biased competition in semantic representation during natural visual search," *NeuroImage*, vol. 216, 2020, Art. no. 116383.
- [54] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *Neuroimage*, vol. 45, no. 1, pp. S199–S209, 2009.
- [55] S.-Y. Xie et al., "Brain fMRI processing and classification based on combination of PCA and SVM," in *Proc. Int. Joint Conf. Neural Netw.*, 2009, pp. 3384–3389.
- [56] C.-Y. Wee et al., "Identification of MCI individuals using structural and functional connectivity networks," *NeuroImage*, vol. 59, no. 3, pp. 2045–2056, 2012.
- [57] S. M. Smith et al., "Functional connectomics from resting-state fMRI," *Trend. Cogn. Sci.*, vol. 17, no. 12, pp. 666–682, 2013.
- [58] F. Zhao et al., "Diagnosis of autism spectrum disorders using multi-level high-order functional networks derived from resting-state functional MRI," *Front. Hum. Neurosci.*, vol. 12, 2018, Art. no. 184.
- [59] A. Riaz et al., "DeepFMRI: End-to-end deep learning for functional connectivity and classification of ADHD using fMRI," *J. Neurosci. Methods*, vol. 335, 2020, Art. no. 108506.
- [60] L.-L. Zeng et al., "Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI," *EBioMedicine*, vol. 30, pp. 74–85, 2018.
- [61] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
- [62] X. Liu et al., "Distributed graph summarization," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2014, pp. 799–808.
- [63] 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, Aug. 2011.
- [64] P. Han et al., "GCN-MF: Disease-gene association identification by graph convolutional networks and matrix factorization," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 705–713.
- [65] Ü. Sakoğlu et al., "A method for evaluating dynamic functional network connectivity and task-modulation: Application to schizophrenia," *Magn. Reson. Mater. Phys. Biol. Med.*, vol. 23, no. 5, pp. 351–366, 2010.
- [66] E. A. Allen et al., "Tracking whole-brain connectivity dynamics in the resting state," *Cereb. Cortex*, vol. 24, no. 3, pp. 663–676, 2014.
- [67] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1310–1318.
- [68] A. A. Ismail et al., "Input-cell attention reduces vanishing saliency of recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, Art. no. 970.
- [69] Y. Zhang et al., "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.
- [70] H. A. Bedel, I. Şivgin, and T. Çukur, "A graphical network layer for lagged analysis of fMRI data," in *Proc. Signal Process. Commun. Appl. Conf.*, 2022, pp. 1–4.
- [71] O. Dalmaz, M. Yurt, and T. Çukur, "ResViT: Residual vision transformers for multimodal medical image synthesis," *IEEE Trans. Med. Imag.*, vol. 41, no. 10, pp. 2598–2614, Oct. 2022.
- [72] A. Schaefer et al., "Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI," *Cereb. Cortex*, vol. 28, no. 9, pp. 3095–3114, 2018.
- [73] Y. Gao and A. No, "Age estimation from fMRI data using recurrent neural network," *Appl. Sci.*, vol. 12, no. 2, 2022, Art. no. 749.
- [74] S. Arslan et al., "Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity," in *Proc. Graph Biomed. Image Anal. Integrating Med. Imag. Non-Imag. Modalities*, 2018, pp. 3–13.
- [75] D. Miwa, V. N. L. Duy, and I. Takeuchi, "Valid P-value for deep learning-driven salient region," 2023, *arXiv:2301.02437*.
- [76] T. Yarkoni et al., "Large-scale automated synthesis of human functional neuroimaging data," *Nat. Methods*, vol. 8, no. 8, pp. 665–670, 2011.
- [77] E. A. Allen et al., "A baseline for the multivariate comparison of resting-state networks," *Front. Syst. Neurosci.*, vol. 5, 2011, Art. no. 2.
- [78] R. A. Poldrack and T. Yarkoni, "From brain maps to cognitive ontologies: Informatics and the search for mental structure," *Annu. Rev. Psychol.*, vol. 67, pp. 587–612, 2016.
- [79] 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.
- [80] M. D. Lieberman et al., "Social, self, (situational), and affective processes in medial prefrontal cortex (MPFC): Causal, multivariate, and reverse inference evidence," *Neurosci. Biobehavioral Rev.*, vol. 99, pp. 311–328, 2019.
- [81] F. Van Overwalle et al., "A functional atlas of the cerebellum based on NeuroSynth task coordinates," *Cerebellum*, vol. 23, pp. 993–1012, 2024.
- [82] M. Shahdloo et al., "Task-dependent warping of semantic representations during search for visual action categories," *J. Neurosci.*, vol. 42, no. 35, pp. 6782–6799, 2022.
- [83] Y. Korkmaz, S. U. H. Dar, M. Yurt, M. Özbey, and T. Çukur, "Unsupervised MRI reconstruction via zero-shot learned adversarial transformers," *IEEE Trans. Med. Imag.*, vol. 41, no. 7, pp. 1747–1763, Jul. 2022.
- [84] A. Güngör et al., "Adaptive diffusion priors for accelerated MRI reconstruction," *Med. Image Anal.*, vol. 88, 2023, Art. no. 102872.
- [85] M. Özbey et al., "Unsupervised medical image translation with adversarial diffusion models," *IEEE Trans. Med. Imag.*, vol. 42, no. 12, pp. 3524–3539, Dec. 2023.
- [86] R. Pope et al., "Efficiently scaling transformer inference," 2022, *arXiv:2211.05102*.
- [87] G. Flandin et al., "Improved detection sensitivity in functional MRI data using a brain parcelling technique," in *Proc. Int. Med. Image Comput. Comput. Assist. Interv.*, 2002, pp. 467–474.
- [88] G. Elmas et al., "Federated learning of generative image priors for MRI reconstruction," *IEEE Trans. Med. Imag.*, vol. 42, no. 7, pp. 1996–2009, Jul. 2023.
- [89] X. Li et al., "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med. Image Anal.*, vol. 65, 2020, Art. no. 101765.