12-12-2022

Spatio-temporal Deep Learning Architectures for Data-Driven Learning of Brain's Network Connectivity

Usman Mahmood

Spatio-temporal Deep Learning Architectures for Data-Driven Learning of Brain's Network Connectivity

by

Usman Mahmood

Under the Direction of Sergey Plis, Ph.D.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2022

ABSTRACT

Brain disorders are often linked to disruptions in the dynamics of the brain's intrinsic functional networks. It is crucial to identify these networks and determine disruptions in their interactions to classify, understand, and possibly cure brain disorders. Brain's network interactions are commonly assessed via functional (network) connectivity, captured as an undirected matrix of Pearson correlation coefficients. Functional connectivity can represent static and dynamic relations. However, often these are modeled using a fixed choice for the data window. Alternatively, deep learning models may flexibly learn various representations from the same data based on the model architecture and the training task. The representations produced by deep learning models are often difficult to interpret and require additional posthoc methods, e.g., saliency maps. Also, deep learning models typically require many input samples to learn features and perform the downstream task well. This dissertation introduces deep learning architectures that work on functional MRI data to estimate disorder-specific brain network connectivity and provide high classification accuracy in discriminating controls and patients. To handle the relatively low number of labeled subjects in the field of neuroimaging, this research proposes deep learning architectures that leverage self-supervised pre-training to increase downstream classification. To increase the interpretability and avoid using a posthoc method, deep learning architectures are proposed that expose a directed graph layer representing the model's learning about relevant brain connectivity. The proposed models estimate task-specific directed connectivity matrices for each subject using the same data but training different models on their own discriminative tasks. The proposed architectures are tested with multiple neuroimaging datasets to discriminate controls and patients with schizophrenia, autism, and dementia, as well as age and gender prediction. The proposed approach reveals that differences in connectivity among sensorimotor networks relative to default-mode networks are an essential indicator of dementia and gender. Dysconnectivity between networks, especially sensorimotor and visual, is linked with schizophrenic patients. However, schizophrenic patients show increased intra-network default-mode connectivity compared to healthy controls. Sensorimotor connectivity is vital for both dementia and schizophrenia prediction, but the differences are in inter and intra-network connectivity.

INDEX WORDS: Self-supervised Learning, Dynamic directed connectivity, Interpretable deep learning, resting state fMRI, brain disorders

Spatio-temporal Deep Learning Architectures for Data-Driven Learning of Brain's Network Connectivity

by

Usman Mahmood

Committee Chair:	Sergey Plis
Committee:	Vince Calhoun
	Rolando Estrada
	Daniel Takabi

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2022

DEDICATION

To my loving wife Araf, for her support, her sacrifices, her belief in my abilities, and always lifting me when I was down. You were my biggest support throughout this journey. This would have been much harder without you. To my parents, for giving me a good platform and working hard on my early education. To my mother, who always wanted me to be a doctor, this is the best doctor I could be. To my brother, whom I followed in many things and who always thought highly of me.

ACKNOWLEDGMENTS

I thank my research advisor, Dr. Sergey Plis, for his guidance, feedback, meetings, and countless discussions. For helping with technical topics and steering me to be a much better researcher. I am also grateful for his efforts toward my future goals and for helping me make valuable connections. I also want to thank my dissertation committee members, Dr. Vince Calhoun, Dr. Rolando Estrada, and Dr. Daniel Takabi, for their valuable feedback and suggestions. I want to thank Dr. Sergey Plis, Dr. Vince Calhoun, and Dr. Satrajit Ghosh for their help in writing and editing this dissertation. I want to thank Dr. Zening Fu for their help in collecting and pre-processing the datasets used in this dissertation and for helping to explain the pre-processing. I thank Mr. Mahfuzur Rahman, Mr. Alex Fedorov, and Mr. Noah Lewis for contributing to the work shown in Chapter 3. Last but not least, I would like to thank my master's professor and research advisor Dr. Khawaja Suleman, for seeing my potential in research and pushing me toward a doctoral.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Brain disorders are often driven by disruptions in the dynamics of the brain's intrinsic functional networks, making it extremely important to identify these networks and determine disruptions in their dynamics. For example, (Culbreth et al. 2021; Yu et al. 2011; Zhang et al. 2019; Zhu et al. 2020; Morgan et al. 2020a; Lynall et al. 2010; van den Heuvel et al. 2010) show that schizophrenic patients have high level of functional disconnectivity between brain networks. Dysregulated brain dynamics and dysregulated dynamic connectivity across the brain's multiple functional networks is seen in Schizophrenic patients (Supekar et al. 2019). In Alzheimer's disease (AD), disrupted brain dynamics demonstrate cognitive dysfunction (Haan et al. 2011). (Cordova-Palomera et al. 2017) suggests that brains of AD patients display altered oscillatory patterns and functional coupling alterations along with decreased global metastability. Alterations in brain activity have been linked to autism spectral disorder (ASD), (Just et al. 2012; Yahata et al. 2016) show dysfunctional brain activity among brain's functional network for ASD patients. (Zeng et al. 2017) show significantly lower whole-brain activity for the ASD group.

It is possible to indirectly measure brain function activity to various degrees of precision using brain imaging methods, such as functional magnetic resonance imagining (fMRI). fMRI captures the nuances of spatio-temporal dynamics that could potentially provide clues to the causes of mental disorders and enable early diagnosis. However, the obtained data for a single subject is of high dimensionality (often in thousands) m and to be useful for learning, and statistical analysis, one needs to collect datasets with a large number of subjects n. Yet, for any kind of a disorder, demographics or other types of conditions, a single study is rarely able to amass datasets large enough to go out of the $m \gg n$ mode.

fMRI captures voxel-level data and does not provide intrinsic functional networks nor their connectivity. It is extremely challenging for any machine learning (ML) or deep learning (DL) model to work directly on the voxel-level data to even perform classification between patients and healthy controls (HC). Therefore, to reduce the number of features, methods are used to get regions made up of several voxels and the connectivity between these regions. To spatially split the brain into networks, existing studies either divide the brain into multiple regions using existing pre-defined brain atlases such as Shaefer (Schaefer et al. 2017), and many others, or estimate constituent components using inference methods, such as independent component analysis (ICA) (Hyvärinen & Oja 2000). Whereas, the connectivity is often assessed via the functional (network) connectivity (F(N)C). Although any statistical dependence measure can be used to represent the FC or FNC, almost always FC or FNC is represented as an undirected correlation matrix of Pearson correlation coefficient (PCC) between the regions/components.

These hand-crafted features (correlation matrices) are used in studies of the brain have demonstrated the overarching value of inspecting the brain and its disorders through the undirected weighted graph of the fMRI correlation matrix. (Yan et al. 2017) uses FC as features to predict schizophrenia-related changes. Whereas, (Parisot et al. 2018) uses FC alongside phenotypic and imaging data as inputs to extract graph features for the classification of AD and Autism. (Kawahara et al. 2016) uses connection strength between brain regions as edges, typically defined as the number of white-matter tracts connecting the regions. (Ktena et al. 2017) employs spectral graph theory to learn similarity metrics among functional connectivity networks.

ML and DL methods can use FC matrices to perform classification between patients and HC with high accuracy. Many studies use FC to predict the gender or disease/disorder (Arslan et al. 2018; Kazi et al. 2021; Kim & Ye 2020; Ktena et al. 2018; Ma et al. 2019) using graph neural networks (GNNs) or other such methods. However, the dynamics of brain function vanishes into proxy features such as correlation matrices of FC. Correlation based FC matrices have many shortcomings including but not limited to inflexibility in terms of the downstream task, undirected relations among regions and networks, and limitation in capturing temporal dynamics.

One of the aims of this research is to show that dynamic DL architectures can be created that work directly on the BOLD time courses and learn task-dependent directed connectivity structures between networks for individual subjects. Interpretation of these estimated connectivity structures could lead to useful insights regarding brain functionality and multiple brain disorders.

This dissertation presents these DL architectures and shows that DL architectures without using hand-crafted features can beat SOTA ML and DL methods that use hand-crafted features in discriminating controls and patients with schizophrenia, autism, and dementia, as well as age and gender prediction from functional MRI data. More importantly, this work shows that connectivity matrices estimated by our DL architectures are more interpretable, are robust to confounding factors, show direction of connectivity between networks and capture more temporal dynamic states than correlation based FC matrices.

1.1 Background and Related Work

This section gives background to the data used in neurogimaging, and also explains why deep learning methods were preferred instead of classical machine learning methods. Current limitations of deep learning methods, especially in the field of neuroimaging are also discussed.

1.1.1 fMRI, Brain Parcellation and Connectivity

fMRI measures BOLD (Blood Oxygenation Level-Dependent) signal that relies on regional differences in cerebral blood flow to delineate regional activity and captures the functional activity of the brain over time with high spatial resolution. fMRI measures the small changes in blood flow that occur during brain activity. Blood flow to the brain is highly locally controlled by the oxygen and carbon dioxide level in the regions of cortex. When brain activity is increased in a region of the cortex, oxygen is extracted from the local capillaries which results into a decrease in local oxygenated hemoglobin, and an increase in locally deoxygenated hemoglobin. In response to this, with a lag of few seconds, cereberal blood flow increases causing a surplus of oxygenated hemoglobin to the region of activity. This oxygenation process is measured during fMRI as deoxygenated hemoglobin is paramagnetic whereas oxygenated hemoglobin is not.

1.1.1.1 Atlas based regions

fMRI measures voxels, but usually relevant brain regions are much larger; therefore, there are different methods to divide a brain, based on structural or functional features into multiple ROIs, where each ROI is a collection of multiple voxels. ROIs help focus on regions rather than individual voxels and help to reduce the number of dimensions as in most studies, voxels are summed/averaged inside a region. Many pre-defined atlases exist (Schaefer et al. 2017; Tzourio-Mazoyer et al. 2002; Desikan et al. 2006) that divide the brain into ROIs based on different techniques like local gradient approach, internal coherence, global similarity, etc.

1.1.1.2 ICA

ICA (Comon 1994; Hyvärinen & Oja 2000) is a computational method to separate a multivariate signal into maximally independent sub-components/ variables. The ability of ICA to extract maximally independent components is instrumental when data is collected via methods like fMRI, which presumably captures the mixture of underlying components of brain activity, it is beneficial to extract the original components using these mixtures. ICA works on two assumptions: 1) The original components are independent, and 2) The components have non-gaussian distribution.

The first assumption is intuitive as ICA finds independent components, giving the minimum number of components required to get the observed data. The basis of the second assumption is the central limit theorem (CLT) that states that the distribution of the sum of two random variables will be more Gaussian than either individual variable, even if the individual variable is non-gaussian itself. ICA uses CLT and the non-gaussian assumption to uncover non-gaussian independent components from the observed data. Mathematically let's assume two variables x_1 and x_2 are observed which are linear combination of the two non-gaussian independent variables s_1 and s_2. Thus;

$$x_1 = a_{11} * s_1 + a_{12} * s_2 \tag{1.1}$$

$$x_2 = a_{21} * s_2 + a_{21} * s_2 \tag{1.2}$$

and let;

$$y_1 = w_{11} * x_1 + w_{12} * x_2 \tag{1.3}$$

$$y_2 = w_{21} * x_2 + w_{21} * x_2 \tag{1.4}$$

According to CLT, y is more gaussian than the signals s as it is a linear combination of them and will be least gaussian when it is directly proportional to either of the independent components. Therefore, increasing the non-gaussainity of y will uncover the original component/variable s. The non-gaussainity is measured by kurtosis (k), as $k = 0$ for the gaussian distribution. Thus ICA is framed as an optimization problem where

$$max \;\; kurt(w1x1 + w2x2) \tag{1.5}$$

$$s = w1 * x1 + w2 * x2 \tag{1.6}$$

1.1.1.3 Correlation-based Functional Connectivity

FC is used in many studies to study brain disorders. In almost all cases FC is computed using PCC which is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{1.7}$$

where $\overline{x}$ and $\overline{y}$ represent mean values. PCC measures the linear correlation between two data samples and holds the commutative property. This study refers to matrices capturing functional connectivity between networks at a whole-brain level as functional network connectivity (FNC) (Jafri et al. 2008; Allen et al. 2011b) and when operating on ROIs – as FC.

1.1.2 Deep Learning

This section explains the pros and cons of machine learning (ML) and deep learning (DL) methods the reason of choosing DL over classical ML methods such as support vector machine (SVM), logistic regression (LR), and many others.

1.1.2.1 Why DL?

Classical machine learning algorithms (SVM, LR) have proven to be highly efficient in classification tasks (Williams et al. 2006; Jalil et al. 2010). These algorithms work on hand-crafted features, e.g., FC matrices (Yan et al. 2017), and produce state-of-the-art results (Douglas et al. 2011). Modern biomedical imaging, functional MRI, collects high dimensional data, where the number of measured values (m) per sample can exceed tens of thousands. Machine learning algorithms do not provide good classification performance in this case as the data dimensions are much higher than the number of data samples (n) available for training, thus creating the curse of dimensionality ($m \gg n$). Dimensionality reduction methods, such as FC, reduce these dimensions and provide hand-crafted features. These hand-crafted features are then necessarily void of many valuable properties and dynamics initially present in the data. On the other hand, using data dynamics is essential for finding distinct features and their relations, responsible for classifying and understanding the system/brain. These dynamics are crucial for the neuroimaging field, where causes and functional/effective connectivity between brain regions of the underlying brain disorders are still unclear. Finding the causes of disorders and the underlying brain networks' connectivity can potentially help prevent, delay and even cure these disorders. As DL methods have the ability to work on raw data and extract task-based useful features, an argument can be made that DL can provide us with useful insights in the field of neuroimaging. DL has its own challenges, one of them is the requirement of a lot of labelled subjects, a problem which is typically solved using self-supervised pre-training which are discussed and applied in the models proposed in this research.

1.1.2.2 Pre-Training

In many fields (Hénaff et al. 2019; Devlin et al. 2018; Lugosch et al. 2019) researchers traditionally employ un-supervised/self-supervised pre-training (Erhan et al. 2010). Furthermore, self-supervised methods with mutual information objective can perform competitively with supervised methods (Oord et al. 2018; Hjelm et al. 2018b; Bachman et al. 2019) and are suitable for several applications (Anand et al. 2019; Ravanelli & Bengio 2018). Pre-training and transfer learning have also been used for neuro/brain imaging (Mensch et al. 2017; Thomas et al. 2019; Mahmood et al. 2020a; Li et al. 2018). Self/un-supervised pre-training on large unlabeled data helps the model to learn useful embeddings/representations, which are then transferred over during training on the small labeled dataset for the downstream task. Pre-training to a certain level can bypass the need to acquire large labeled training datasets and achieve higher classification performance than non-pre-trained counterparts (Mahmood et al. 2021a).

1.1.2.3 Self-attention

The concept of self-attention has been applied to a variety of tasks related to NLP such as; reading comprehension, abstractive summarization, and text representations (Cheng et al. 2016; Lin et al. 2017; Parikh et al. 2016; Paulus et al. 2017). Self-attention allows the inputs to interact among themselves and find out global dependencies among the inputs and outputs. The dependencies are represented as weights, which measure the attention given to inputs by other inputs. The outputs given by the module are linear combinations of the inputs and the relative weights (dependencies/attention). The critical advantage of self-attention is that it is not commutative, meaning the attention given by two inputs to each other does not have to be equal. Self-attention is most commonly used for inputs of a sequence through time, but this work uses the concept to find dependencies between regions/components/sequences with a goal creating an attention module that learns connectivity between brain regions.

The self-attention model creates three embeddings namely (key, query, value) for each input, which are usually created using simple linear layers to map the inputs to a smaller dimension. With ϕ representing a linear layer, $key_i = \phi_k(input_i)$, $query_i = \phi_q(input_i)$, $value_i = \phi_v(input_i)$. To create weights between an input and every other input, the model takes dot product of an inputs's query with every other input's key embedding to get scores between them. Hence, $score_{ij} = query_i \cdot key_j$. The scores are then converted to weights using softmax. $w_i = Softmax(score_i)$ where $score_i \in \mathbb{R}^{1\times r}$ is a vector of scores between region i and every other region. The weights are then multiplied with the *value* embedding of each input and summed together to create a new representation for $input_i$. The following equations show how to get new input embedding and weight values.

$$\begin{aligned}
&key_i = input_i * W^{(k)}, \;\; value_i = input_i * W^{(v)}, \;\; query_i = input_i * W^{(q)} \\
&K = ||_{i=1}^{r} key_i^T = key_i^T||....||key_r^T, \;\; weight_i = softmax(query_i * K) \qquad (1.8) \\
&new\_input_i = \sum_{j}^{r}(weight_{ij} * value_j)
\end{aligned}$$

This process is carried out for all the regions, producing a new representation of every input and the weights between inputs.

1.1.2.4 Graph Neural Networks

Datasets from different fields are represented as a graph. Graph networks (Scarselli et al. 2009; Bruna et al. 2014) are proposed to work on such datasets. Recently, GNN (Graph Neural Networks) have been extensively used to learn representations on graph-structured data (Bronstein et al. 2017; Hamilton et al. 2018; Gilmer et al. 2017; Parisot et al. 2018). GNNs take nodes from data and update representations of nodes with the help of different aggregating functions. The aggregate functions work using a message-passing system, where a node receives messages from its neighbors, which are defined by edges. Lets represent a graph G with V, A, E where $V \in \mathbb{R}^{n \times m}$ is the matrix of vertices, with x^{n_i} representing the i^{th} node/vertex, having m dimensions. $A, E \in \mathbb{R}^{n \times n}$ are the adjacency and edge weight matrices. A GNN module takes multiple steps, where at every step s, each node aggregates feature vectors of every other node relative to the weight edge between the nodes and pass the resultant, and its own feature vector through another neural network to obtain new embedding for itself.

$$x_s^{n_i} = \phi(x_{s-1}^{n_i}, \bigcup_{\forall n_j : n_j -> n_i} e_{ji} x_{s-1}^{n_j}) \tag{1.9}$$

Here $\bigcup$ is the aggregate function which can be sum, max, average, or any other function which is permutation invariant. ϕ represents any neural network such as GRU. The learned representations can then be used for node classification, graph classification, or predicting edges between nodes by using an existing true graph structure or learning the graph (Monti et al. 2016; Velickovic et al. 2018; Kipf & Welling 2017; Gilmer et al. 2017; Zitnik et al. 2018; Zhang & Chen 2018; Wang et al. 2019; Kipf et al. 2018; Zitnik et al. 2018).

1.1.3 Challenges and Limitations

The association of brain disorders with abnormal static or dynamic functional connectivity highlights the need to develop models that can identify disorder-specific connectivity aberrations. This observation guides development of various approaches to brain connectivity analysis (Yan et al. 2017; Parisot et al. 2018; Ktena et al. 2017; Arslan et al. 2018; Kazi et al. 2021; Kim & Ye 2020; Ktena et al. 2018; Ma et al. 2019). However in most existing approaches, the functional connectivity matrices are not informed by the prediction task but instead estimated prior to training; thus, they depend entirely on the chosen input window of data samples. The independence from the downstream task results in inflexible estimation of connectivity matrices as the estimate is unchanged regardless of whether the task is to predict a brain disorder, age, or other quantity. (Kim et al. 2021) proposed a method where the functional connectivity structure is computed based on the learned representations of the data, but even this method lacks a learnable connectivity estimation method. This study makes and argument that task-dependent connectivity matrices can be estimated by a deep learning (DL) model using learnable weights. DL models are flexible in their ability to learn a variety of representations from the same data based on the architecture and ground-truth signal used in training.

However, using a DL method to estimate a connectivity matrix can be challenging without the presence of the ground-truth graph during training. Another problem of many DL models is lack of consistency and interpretability in the learned representations. Saliency maps commonly used to address interpretability of these models (Simonyan et al. 2014; Ras et al. 2021; Angelov et al. 2021; Lewis et al. 2021) may be difficult to interpret (Liu et al. 2021). Arguably, the difficulty of interpreting representations is the reason why studies using DL models incorporate inflexible but interpretable feature selection steps for connectivity estimation, for example Pearson correlation coefficients (PCC) (Freedman et al. 2007).

In most of the current studies, functional connectivity estimates are either static or dynamically computed using a sliding window approach dependent on the window size and stride (Fu et al. 2018; Damaraju et al. 2014; Armstrong et al. 2016; Gadgil et al. 2021; Yao et al. 2020; Fu et al. 2020). Unable to capture non-stationarity, static matrices miss essential information about dynamics. For example, dynamic functional connectivity estimates show re-occurring patterns which cannot be captured by their static counterparts (Allen et al. 2012; Hutchison et al. 2013; Calhoun et al. 2014). Using a static graph learning method to capture a dynamical system may reduce classification performance (da Xu et al. 2020). (Kipf et al. 2018) show improved results by just dynamically re-evaluating the learned static graph during testing. The improved performance for the relevant task is understandable as the dynamic connectivity provides essential information about the system, for instance, capturing re-occurring patterns. The brain's functional activity is also perceived to be highly dynamic and hence cannot be faithfully captured with a static or even window-based approach (Yaesoubi et al. 2018).

Furthermore, studies using functional connectivity to measure connectivity between brain regions or networks do not capture the direction of interaction and only measure undirected statistical dependence such as correlations, coherence, or transfer entropy. Correlation can arise for many reasons; for example, due to a common cause when an unobserved network affects two networks that are observed (Spirtes et al. 1993; Pearl 2000). Arguably, dynamics of interaction among brain networks is beyond simple correlations and correlation may only partially describe it. Whereas, effective connectivity is a more general way to represent dynamic and directed relationships among brain's intrinsic networks. As introduced by

(Friston 2011) effective connectivity falls into a model-based class of methods while multiple other methods, including those in the model-free class have been since developed (Chickering 2002a; Spirtes & Glymour 1991; Chickering 2002b; Bielza & Larranaga 2014; Goebel et al. 2003; Deshpande et al. 2011; Mitra et al. 2014; Seth et al. 2015; Schreiber 2000; Vicente et al. 2011; Ursino et al. 2020; Gorrostieta et al. 2013; Chiang et al. 2017).

1.2 Contributions

In this dissertation work, the following objectives were made and accomplished:

1. Develop novel pre-training/transfer learning methods to increase the classification performance of deep learning models in neuroimaging, where $m >> n$. See Chapter: 3.

2. Construct models that automatically identify disease-specific brain networks specified based on existing brain atlases. See Chapter: 4.

3. Build deep learning methods that, in an end-to-end manner for each subject, estimate connectivity structures between the disease-specific spatially mapped brain networks. See Chapter: 4 and 5.

4. Develop deep learning architecture which produces dynamic, interpretable and directed connectivity matrices without the help of an additional posthoc interpretability method. Introspection of the estimated graphs and discover disorder specific spatial and temporal bio-markers for multiple brain disorders. See Chapter: 5.

CHAPTER 2
INPUT DATA

This section presents the different datasets used throughout this study and the different pre-processing piplelines used.

2.1 fMRI Data

In this work, resting state functional magnetic resonance imaging (rs-fMRI) data as input to our models. Six brainimaging datasets used in this study are collected from FBIRN (Function Biomedical Informatics Research Network [1]) (Keator et al. 2016) project, from COBRE (Center of Biomedical Research Excellence) (Çetin et al. 2014) project, from release 1.0 of ABIDE (Autism Brain Imaging Data Exchange [2]) (Di Martino et al. 2014) and from release 3.0 of OASIS (Open Access Series of Imaging Studies [3]) (Rubin et al. 1998). Healthy controls from the HCP (Human Connectome Project [4]) (Van Essen et al. 2013) are used for gender prediction. Subjects from ABCD (Adolescent Brain Cognitive Development [5]) (Casey et al. 2018) are also used for gender prediction. Refer to Table 2.1 for details of the datasets. Datasets used for specific model is mentioned in relevant chapters.

2.1.1 Preprocessing

Two typical brain parcellation techniques are used in this work; independent component analysis (ICA) and regions of interest (ROIs) based on a pre-defined atlas. The preprocessing

[1] FBIRN phase III is used.

[2] `http://fcon_1000.projects.nitrc.org/indi/abide/`

[3] `https://www.oasis-brains.org/`

[4] Scans from the first session are used.

[5] First scans from the first session are used.

pipeline used depends on the parcellation technique and the pipeline used in state-of-the-art studies for the dataset. All the preprocessing was done before training the model.

2.1.1.1 ICA Parcellation:

For all experiments conducted using ICA as brain parcellation technique the fMRI data was preprocessed using statistical parametric mapping (SPM12, `http://www.fil.ion.ucl.ac.uk/spm/`) under the MATLAB 2021 environment. A rigid body motion correction was performed to correct subject head motion, followed by the slice-timing correction to account for timing difference in slice acquisition. The fMRI data were subsequently warped into the standard Montreal Neurological Institute (MNI) space using an echo planar imaging (EPI) template and were slightly resampled to $3 \times 3 \times 3$ mm^3 isotropic voxels. The resampled fMRI images were then smoothed using a Gaussian kernel with a full width at half maximum (FWHM) = 6 mm.

Subjects are selected for further analysis (Fu et al. 2021a) if the subjects have head motion $\leq 3°$ and ≤ 3 mm, and with functional data providing near full brain successful normalization (Fu et al. 2019). 100 ICA components are estimated using a novel fully automated Neuromark pipeline "neuromark_fmri_1.0"[6] described in (Fu et al. 2019). This method is capable of capturing robust imaging features that are comparable across subjects, datasets, and studies, which is beneficial for those studies need replication. The Neuromark framework leverages an adaptive-ICA technique that automates the estimation of comparable brain markers across subjects, datasets, and studies. A set of component templates were

[6] `https://trendscenter.org/data/`

used as references to guide the estimation of single-scan components for the data. These component templates were created via a unified ICA pipeline. They were constructed using an independent resting-state fMRI data with large samples of healthy subjects from the genomics superstruct project (GSP). The GSP data include 1005 subjects' scans that passed the data QC. High model order (order = 100) group ICA was performed on the GSP data, and then the independent components (ICs) from the GSP data were used as the references to extract components for each dataset used for experiment in this study. The Neuromark framework extracts the components for each subject respectively, which means that the estimation of features of each subject is not influenced by the others. However, the choice of components (and number of components) can influence accuracy, but our study is not focusing on determining the best number of ICs rather use the available components and let the model decide the task-dependant components.

Table 2.1: Details of the datasets used throughout this research.

Name	Category	Preprocessing	Parcellation	Subjects	0 Class	1 Class	time-points
FBIRN	Schizophrenia	SPM12	ICA	311	151	160	157
OASIS	Dementia	SPM12	ICA	912	651	261	157
ABIDE	Autism	SPM12	ICA	569 (TR=2)	255	314	140
ABIDE	Autism	SPM12	ICA	869	398	471	140
HCP	Gender	SPM12	ICA	833	390	443	980
ABCD	Gender	SPM12	ICA	10976	5697	5279	370
FBIRN	Schizophrenia	SPM12	Shaefer 200	311	151	160	157
HCP	Gender	Glasser	Shaeffer 200	942	411	531	1200
ABIDE	Autism	C-PAC	Shaeffer 200	871	403	468	83-316

2.1.1.2 Region Parcellation:

State-of-the-art methods use different preprocessing pipelines for different datasets. For comparison with these methods on HCP, ABIDE, and FBIRN datasets, the same preprocessing pipelines as in the relevant comparing method were selected. HCP (Van Essen et al. 2013) data used in this study was first minimally pre-processed following the pipeline described in (Glasser et al. 2013). The preprocessing includes gradient distortion correction, motion correction, and field map preprocessing, followed by registration to T1 weighted image. The registered EPI image was then normalized to the standard MNI152 space. To reduce noise from the data, FIX-ICA based denoising was applied (Salimi-Khorshidi et al. 2014; Griffanti et al. 2014). To minimize the effects of head motion subject scans with framewise displacement (FD) over 0.3mm at any time of the scan were discarded. The FD was computed with fsl motion outliers function of the FSL (Jenkinson et al. 2012). There were 152 discarded scans from filtering out with the FD, and 942 scans were left. For all experiments, the scans from the first run of HCP subjects released under S1200 were used. ABIDE (Di Martino et al. 2014) was pre-processed using C-PAC (Aertsen & Preissl 1991). The preprocessing includes; slice time correction, motion correction, skull striping, global mean intensity normalization, nuisance signal regression, band pass filtering, and finally functional images were registered to anatomical space (MNI12). After pre-processing using C-PAC, 871 out of 1112 subjects were chosen based on the visual quality, inspected by three human experts which looked for brain coverage, high movement peaks and other artifacts resulted by scanner (Abraham et al. 2017; Parisot et al. 2018; Cao et al. 2021). To pre-process FBIRN data, SPM12 pipeline was used as explained in previous section with few extra steps. After the smoothing using a Gaussian kernel, the functional images were temporally filtered by a finite impulse response (FIR) bandpass filter (0.01 Hz-0.15 Hz). Then for each voxel, six rigid body head motion parameters, white matter (WM) signals, and cerebrospinal fluid (CSF) signals were regressed out using linear regression.

In this work, in total three atlases were used for brain parcellation; Shaefer (Schaefer et al. 2017), automated anatomical labeling (AAL) (Tzourio-Mazoyer et al. 2002), and Harvard Oxford (HO) (Desikan et al. 2006), a with 200, 116, and 111 regions respectively. For each region, average value is computed for all the voxels falling inside a region, thus resulting into a single time-series for each region. After dividing data into regions, each time-series was standardized by their zscore having zero mean and unit variance.

CHAPTER 3

PRE-TRAINING AND SELF-SUPERVISED LEARNING

Un/Self-supervised pre-training is a well-known technique to get a head start for the deep neural network (Erhan et al. 2010). It finds wide use across a number of fields such as computer vision (Hénaff et al. 2019), natural language processing (NLP) (Devlin et al. 2018) and automatic speech recognition (ASR) (Lugosch et al. 2019). However, outside NLP unsupervised pre-training is not as popular as supervised.

Recent advances in self-supervised methods with mutual information objectives are approaching performance of supervised training (Oord et al. 2018; Hjelm et al. 2018b; Bachman et al. 2019) and can scale pre-training to very deep convolutional networks (e.g., 50-layer ResNet). They were shown to benefit structural MRI analysis (Fedorov et al. 2019), learn useful representations from the frames in Atari games (Anand et al. 2019) and for speaker identification (Ravanelli & Bengio 2018). Pre-trained models can outperform supervised methods by a large margin in case of small data (Hénaff et al. 2019).

Earlier work in brain imaging (Khosla et al. 2019b; Plis et al. 2014) have been based on unsupervised methods to learn the dynamics and structure of the brain using approaches such as ICA (Calhoun et al. 2001) and HMM (Eavani et al. 2013). Deep learning for capturing the brain dynamics has also been previously proposed (Hjelm et al. 2014, 2018a; Khosla et al. 2019a). In some very small datasets, transfer learning was proposed for use in neuroimaging applications (Mensch et al. 2017; Li et al. 2018; Thomas et al. 2019). Yet another idea is the data generating approach (Ulloa et al. 2018). ST-DIM (Anand et al. 2019) has been used for pre-training on unrelated data with subsequent use for classification (Mahmood et al. 2019b).

One of the goals of this dissertation is to enable the direct study of brain dynamics in the $m \gg n$ situation. In the case of brain data it, in turn, can enable an analysis of brain function via model introspection. This chapter presents a novel self supervised training schema which reinforces whole sequence mutual information local to context (whole MILC). The whole MILC model shows how one can achieve significant improvement in classification directly from dynamical data on small datasets by taking advantage of publicly available large but unrelated datasets. Research work in this chapter demonstrates that it is possible to train a model in a self-supervised manner on dynamics of healthy control subjects from the Human Connectome Project (HCP) (Van Essen et al. 2013) and apply the pre-trained model to a completely different data collected across multiple sites from healthy controls and patients. This chapter shows that pre-training on dynamics allows the encoder to generalize across a number of datasets and a wide range of disorders: schizophrenia, autism, and Alzheimer's disease. Importantly, it is shown that learnt dynamics generalizes across different data distributions, as the proposed model pre-trained on healthy adults shows improvements in children and elderly.

3.1 Method

This chapter presents MILC as a self-supervised pre-training method. MILC is used to pre-train on large unrelated and unlabelled data to better learn data representation. The learnt representations are then used for classification on downstream tasks adding a simple linear network on top of the pre-training architecture. The fundamental idea of MILC is to establish relationship between windows (a time slice from the entire sequence) and their respective sequences through learning useful signal dynamics. In all of the experiments, encoded rsfMRI ICA time courses is used as the sequences and a consecutive chunk of time points as windows. The model uses the idea to distinguish among sequences (subjects) which proves to be extremely useful in downstream tasks e.g classification of HC or SZ subjects. To realize the concept, mutual information of the latent space of a window and the corresponding sequence as a whole is maximized.

Let $D = \{(u_t^i, v^j) : 1 \leq t \leq T, 1 \leq i, j \leq N\}$ be a dataset of pairs computed from ICA time courses. u_t^i is the local embedding of t-th window taken from sequence i, v^j is the global embedding for the entire sequence j. T is the number of windows in a sequence, and N is the total number of sequences. Then $D^+ = \{(u_t^i, v^j) : 1 \leq t \leq T, i = j\}$ is called a dataset of positive pairs and $D^- = \{(u_t^i, v^j) : 1 \leq t \leq T, i \neq j\}$ — of negative pairs. The dataset D^+ refers to a joint distribution and D^- — a marginal distribution of the whole sequence and the window in the latent space. Eventually, the lower bound with InfoNCE estimator (Oord et al. 2018) $\mathcal{I}_f(D^+)$ is defined as:

$$\mathcal{I}(D^+) \geq \mathcal{I}_f(D^+) \triangleq \sum_{i=1}^{N} \sum_{t=1}^{T} \log \frac{\exp f((u_t^i, v^i))}{\sum_{k=1}^{N} \exp f((u_t^i, v^k))}, \tag{3.1}$$

where f is a critic function. Specifically, a separable critic $f(u_t, v_s) = \phi(u_t^i)^\intercal(v^j)$, is used, where ϕ is some embedding function parameterized by neural networks. Such embedding function is used to calculate value of a critic function in same dimensional space from two dimensional inputs. Critic learns an embedding function such that critic assigns higher values for positive pairs compared to negative pairs: $f(D^{+}) \gg f(D^{-})$.

The critic function takes the latent representation of a window and sequence as input. This work defines latent state of window as an output z_t^i produced by the CNN part of MILC, given input from t-th window x_t^i of sequence i. The latent state of sequence as c^j is the global embedding obtained from MILC architecture. Thus the critic function for input pair (x_t^i, x^j)—a window and a sequence—is $f = \phi(z_t^i)^{\intercal}(c^j)$. The loss is InfoNCE with f as $L = I_f$. The scheme of the MILC is shown in Figure 3.1.

3.1.1 Transfer and Supervised Learning

In the downstream task, the representation (output) of the attention model pre-trained using MILC is used as input to a simple binary classifier on top. Refer to section 3.2.1 for further details.

3.2 Experiments

In this section, the performance of the model is shown on both, synthetic and real data. Three different variations of the proposed model are shown to test the advantage of pre-training on large unrelated dataset — 1) FPT (Frozen Pre-Trained): The pre-trained model is not further trained on the dataset of downstream task, 2) UFPT (Unfrozen Pre-Trained): The pre-trained model is further trained on the dataset of downstream task and 3) NPT (Not Pre-trained): The model is not pre-trained at all and only trained on the dataset of

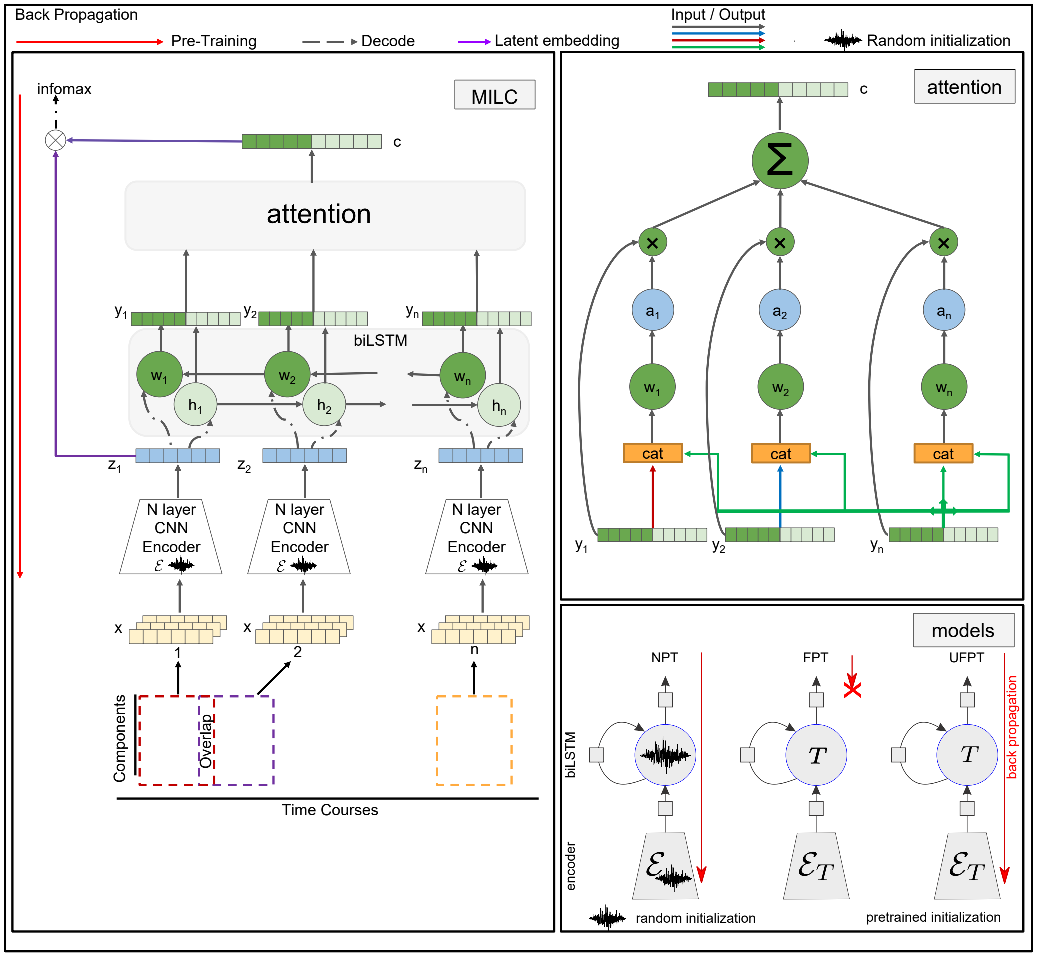


Figure 3.1 **Left:** MILC architecture used in pre-training. ICA time courses are computed from the rsfMRI data. Results contain statistically independent spatial maps (top) and their corresponding time courses. **Right Up:** Detail of attention model used in MILC. **Right Down:** Three different models are used for downstream tasks.

downstream task. The models are shown in Figure 3.1. In each experiment, all three models are compared to demonstrate the effectiveness of unsupervised pre-training.

3.2.1 Setup

The CNN Encoder of MILC for simulation experiment consists of 4 1D convolutional layers with output features $(32, 64, 128, 64)$, kernel sizes $(4, 4, 3, 2)$ respectively, followed by ReLU after each layer followed by a linear layer with 256 units. For real data experiments, the model uses 3 1D convolutional layers with output features $(64, 128, 200)$, kernel sizes $(4, 4, 3)$ respectively, followed by ReLU after each layer followed by a linear layer with 256 units. The model uses stride 1 for all of the convolution layers. Testing is also performed against autoencoder based pre-training for simulation experiment, for which the same CNN encoder is used as for MILC in the reduction phase. For the decoder, the reverse architecture of the encoder is used that results in 10×20 windows at the output.

In MILC based pre-training, for all possible pairs in the batch, feature z from the output layer of CNN encoder is taken. The latent representation of the entire time series is then passed through biLSTM. The output of biLSTM is used as input to the attention model to get a single vector c, which represents the entire time series. Scores are calculated using z and c as explained in 3.1. Loss is computed using these scores. The neural networks are trained using Adam optimizer.

In downstream tasks the higher interest is in subjects for classification task, for each subject the output of attention model (c) is used as input to a feed forward network of two linear layers with 200 and 2 units to perform binary classification. For experiments, a hold out is selected for testing and is never used through the training/validation phase. For each experiment, 10 trials are performed to ensure random selection of training subjects and, in each case, the performance is evaluated on the hold out (test data). The code is available at: `https://github.com/UsmanMahmood27/MILC`

3.2.2 Simulation

To generate synthetic data, multiple 10-node graphs are created with 10×10 stable transition matrices. Using these graphs, multivariate time series are generated with autoregressive (VAR) and structural vector autoregressive (SVAR) models (Lütkepohl 2005).

50 VAR times series with size 10×20000 are split into three time slices respectively for training, validation and testing. Using these samples, MILC is pre-trained to assign windows to respective time series.

In the final downstream task, the model classifies the whole time-series into VAR or SVAR (obtained by randomly dropping 20% VAR samples) groups. 2000 generated samples are split as 1600 for training, 200 for validation and 200 for hold-out test. For both pre-training and downstream task, the same set up as described in section 3.2.1 is followed.

Effectiveness of MILC is compared with the model used in (Mahmood et al. 2019b) and two variations of autoencoder based pre-training. The two variations of autoencoder are acquired by replacing the CNN encoder of (Mahmood et al. 2019b) and MILC by the pre-trained or randomly initialized autoencoder during downstream classification, depending on the model as explained in section 3.2. These two variations are referred as *AE_STDIM* and *AE_STDIM+attention*. Note that difference between the two is the added attention layer in the later during downstream classification.

It is observed that the MILC based pre-trained models can easily be fine-tuned only with small amount of downstream data. Note, with very few samples, models based on the pre-trained MILC (FPT and UFPT) outperform the un-pre-trained models (NPT),

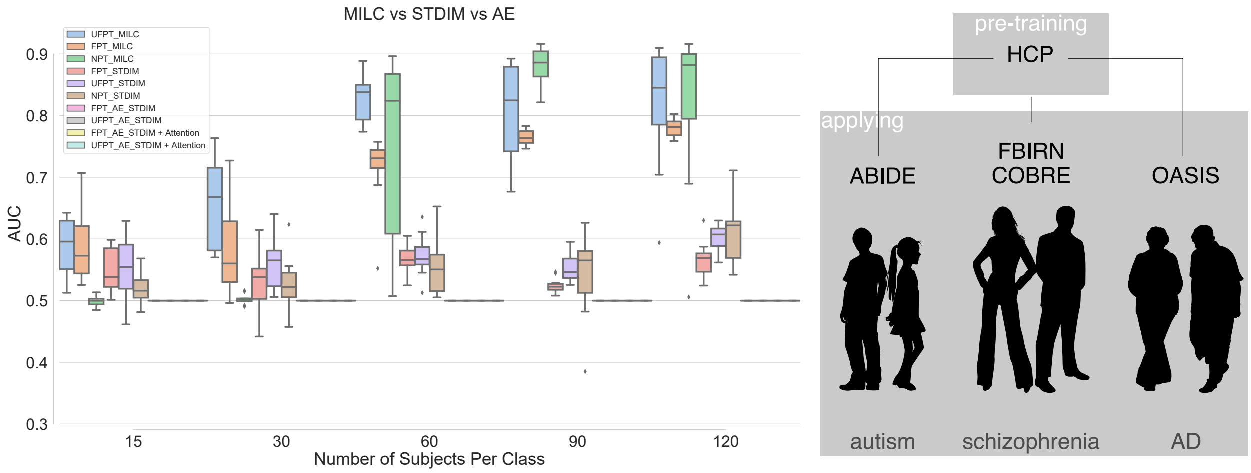


Figure 3.2 **Left:** Area Under Curve (AUC) scores for VAR vs. SVAR time-series classification using MILC, ST-DIM and autoencoder based pre-training methods. MILC based pre-training greatly improves the performance of downstream task with small datasets. On the other side, ST-DIM works better than autoencoder based pre-training which completely fails to learn dynamics and thus exhibits poor performance. **Right:** Datasets used for pre-training and classification tasks. Healthy controls from the HCP (Van Essen et al. 2013) are used for pre-training guided by data dynamics alone[1]. The pre-trained model is then used in downstream classification tasks of 3 different diseases, 4 independently collected datasets, many of which contain data from a number of sites, and consist of populations with significant age difference. The age distributions in the datasets have the following mean and standard deviation: **HCP:** 29.31 ± 3.67; **ABIDE:** 17.04 ± 7.29; **COBRE:** 37.96 ± 12.90; **FBIRN:** 37.87 ± 11.25; **OASIS:** 67.67 ± 8.92.

ST-DIM models, autoencoder based models. ST-DIM based pre-training model (Mahmood et al. 2019b) performs reasonably well compared to autoencoder and NPT models, however, MILC steadily outperforms ST-DIM. Results show that autoencoder based self-supervised pre-training does not assist in VAR vs. SVAR classification. Refer to Figure 3.2 **Left** for the results of simulation experiments.

3.2.3 Brain Imaging

3.2.3.1 Datasets

Next, MILC is applied to brain imagining data. Refer to Figure 3.2 for the details of the datasets used. MILC is compared with ST-DIM based pre-training shown in (Mahmood et al. 2019b).

3.2.3.2 Schizophrenia

For schizophrenia classification, experiments are conducted on two different datasets; FBIRN (Keator et al. 2016) and COBRE (Çetin et al. 2014). The datasets contain labeled Schizophrenia (SZ) and Healthy Control (HC) subjects.

FBIRN

The dataset has total 311 subjects. Two hold-out sets with sizes 32 and 64 are used for validation and test respectively, remaining are used for supervised training. The details of the results are shown in Figure 3.3. We can see that the pre-trained MILC models outperform NPT and also ST-DIM based pre-trained models.

COBRE

The dataset has total 157 subjects — a collection of 68 HC and 89 affected with SZ. Two hold-out sets of size 32 each are used for validation and test respectively. The remaining data is used for supervised training. The results in Figure 3.3 strengthen the efficiency of MILC. That is, with only 15 training subjects, FPT and UFPT perform significantly better

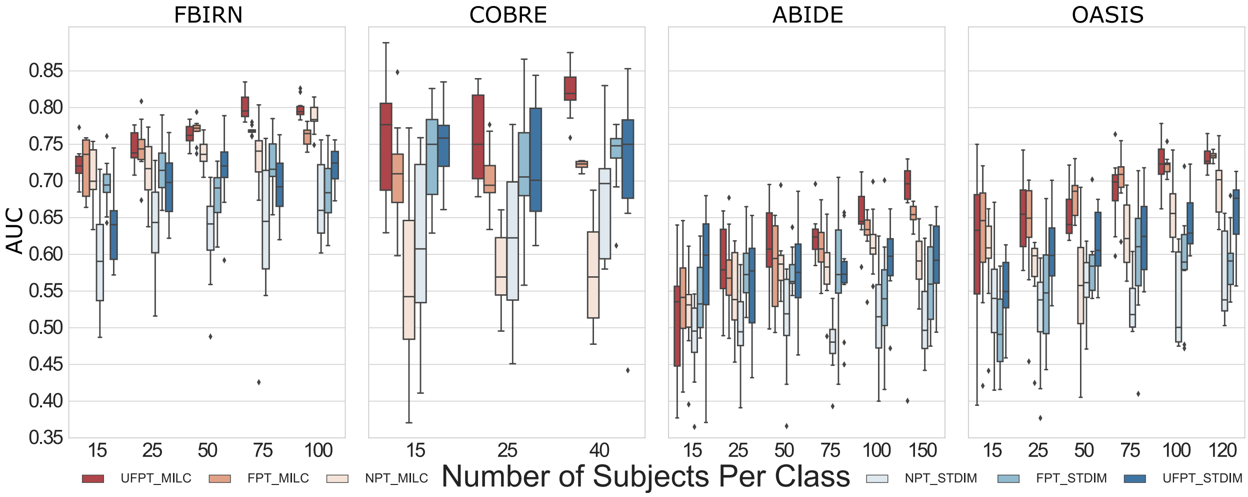


Figure 3.3 AUC scores for all the three models (Refer to Figure 3.1) on real dataset. With every dataset, models pre-trained with MILC (FPT, UFPT) perform noticeably better than not pre-trained model (NPT). Results also show that the learnability of MILC model dramatically increases with small increase in training data (x_axis). As we can see across the datasets, MILC outperforms ST-DIM with a large margin offering ∼ 10% higher AUC when maximum achievable AUC scores are compared.

than NPT having ≃ 0.20 difference in their median AUC scores.

3.2.3.3 Autism

With 569 total subjects, 255 are HC and 314 are affected with autism. 100 subjects are used each for validation and test purpose. The remaining data is used for downstream training i.e., autism vs. HC classification. Figure 3.3 shows, MILC pre-trained models perform reasonably better than NPT and thus reinforces the proposed hypothesis that unsupervised pre-training learns signal dynamics useful for downstream tasks. Possibly, the reason why pre-trained models do not work well for 15 subjects is that the dataset is much different than HCP. The big age gap between subjects of HCP and ABIDE is a major difference and 15 subjects are not enough even for pre-trained models. Refer to Figure 3.2 for the demographic information of all the datasets.

3.2.3.4 Alzheimer's disease

The dataset OASIS (Rubin et al. 1998) has scans for HC and patients suffering from different kind of dementia. For this experiment, only HC and Alzheimer's classification is performed. (186) subjects with Alzheimer's (AD) and equal number of randomly chosen HC are used. Two hold-out sets each of size 64 respectively are used for validation and test purpose. The remaining are used for supervised training. Refer to Figure 3.3 for results. The AUC scores of pre-trained models is higher than NPT starting from 15 subjects, even with 120 subjects NPTdoes not perform equally well.

3.2.4 Saliency

The experiments demonstrate that with the whole MILC pre-training it is possible to achieve reasonable prediction performance from complete dynamics even on small data. Importantly, it is now possible to investigate what in the dynamics was the most discriminative (see Figure 3.4).

3.3 Conclusions

As the work shown in this chapter demonstrates, self-supervised pre-training of a spatio-temporal encoder gives significant improvement on the downstream tasks in brain imaging datasets. Learning dynamics of fMRI helps to improve classification results for all three dieseases and speed up the convergence of the algorithm on small datasets, that otherwise do

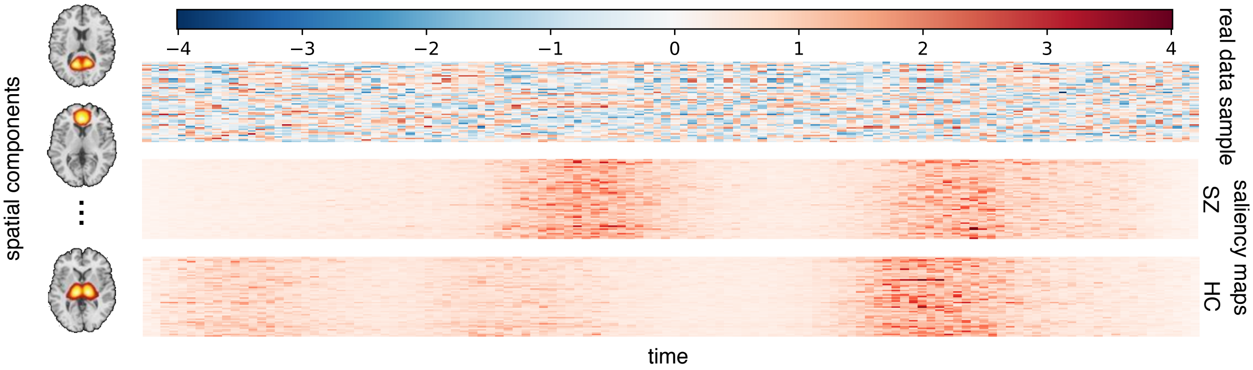


Figure 3.4 Example saliency maps from a pre-trained MILC model: one for a healthy control and one for a schizophrenia subject (FBIRN data). More work is needed, but we can see that not only the proposed model predicts diagnosis but also can point out when during the resting state scan discriminative activity was observed.

not provide reliable generalizations. Although the utility of these results is highly promising by itself, it is expected that further model introspection would yield insight into the spatio-temporal biomarkers of schizophrenia. It may indeed be learning crucial information about dynamics that might contain important clues into the nature of mental disorders. Also, learning disorder-specific graph structure between the input networks can help to discover useful insights.

CHAPTER 4
DISORDER-SPECIFIC GRAPH ESTIMATION

Existing studies often heavily depend on the underlying method of functional connectivity estimation, in terms of classification accuracy, feature extraction, or learning brain dynamics. Studies like (Rashid et al. 2016; Saha et al. 2020; Salman et al. 2019) depend on handcrafted features. These studies work very well on classification but do not learn a sparse graph of brain's network connectivity and not too helpful for identifying bio-markers in the brain. Many functional connectivity studies (Du et al. 2018) on brain disorders utilize ROIs predefined based on anatomical or functional atlases, which are either fixed for all subjects or based are based on group differences.

These approaches ignore the possibility of inter-subject variations of ROIs, especially the variations due to the underlying disease conditions. They also rely on the complete set of these ROIs discounting the possibility that only a small subset may be relevant for the disorder. A disorder can have varying symptoms for different people, hence making it crucial to determine disorder and subject specific ROIs.

This chapter addresses the problems of using a fixed method of learning functional connectivity and using it as a fixed graph to represent brain structure (the standard practices) by utilizing a novel attention based Graph Neural Network (GNN) (Li et al. 2016), called BrainGNN. The proposed model, BrainGNN is applied to fMRI data and 1) achieve comparable classification accuracy to existing algorithms, 2) learn dynamic graph functional connectivity, and 3) increase model interpretability by learning which regions from the set of ROIs are relevant for the classification, enabling additional insights into the health and disordered brain.

4.1 Materials and Methods

4.1.1 Materials

In this chapter, the data from Function Biomedical Informatics Research Network (FBIRN) (Keator et al. 2016) dataset is used to train and test BrainGNN. The dataset includes schizophrenia (SZ) patients and healthy controls (HC). Resting fMRI data from the phase III FBIRN were analyzed for this project. The dataset has 368 total subjects out of which 311 were selected based on the preprocessing method explained in Section 2.1.1.1. To partition the data into regions use automated anatomical labeling (AAL) (Tzourio-Mazoyer et al. 2002) which contains 116 brain regions. Taking sum of the voxels inside a region is an easy and common method but this gives and unfair advantage to bigger regions. For this, the weighted average of the voxel intensities inside a region are taken instead of summation. Weight is the value of a voxel being inside a region, as these values are not binary. Averaging helps to negate the bias towards bigger regions. This results in a dataset $D = (S_1, S_2, S_3......S_n)$ where $S_i \in \mathbb{R}^{r\times \mathrm{t}}$, $n = 311$, $r = 116$, $t = 160$.

4.1.2 Method

The architecture of the proposed model has three distinct parts: 1) a Convolutional Neural Network (CNN) (Lecun et al. 1998) that creates embeddings for each region, 2) a Self-Attention mechanism (Vaswani et al. 2017) that assigns weights between regions for functional connectivity and 3) A GNN that uses regions (nodes) and edges for graph classification. This section explains the purpose and details of each part separately. Refer to Figure 4.1 for the complete architecture diagram of BrainGNN.

4.1.2.1 CNN Encoder

A CNN (Kiranyaz et al. 2021) encoder is used to obtain the representation of individual regions created in the preprocessing step outlined in 4.1.1. Each region vector of dimension t = 160 is passed through multiple layers of one dimensional convolution, and a fully connected layer to get final embedding. The one dimensional CNN encoder used in the architecture consists of 4 convolution layers with filter size $(4, 4, 3, 1)$, stride $(2, 1, 2, 1)$ and output channels $(32, 64, 64, 10)$. This is followed by a fully connected layer resulting in a final embedding of size 64. The model uses rectified linear unit (ReLU) as an activation layer between convolution layers. Each region is encoded individually to later on create connections between regions and interpret which regions are more important/informative for classification. The one dimensional CNN layer embeds the temporal features of regions and the spatial connections are handled in the attention and GNN parts of the architecture.

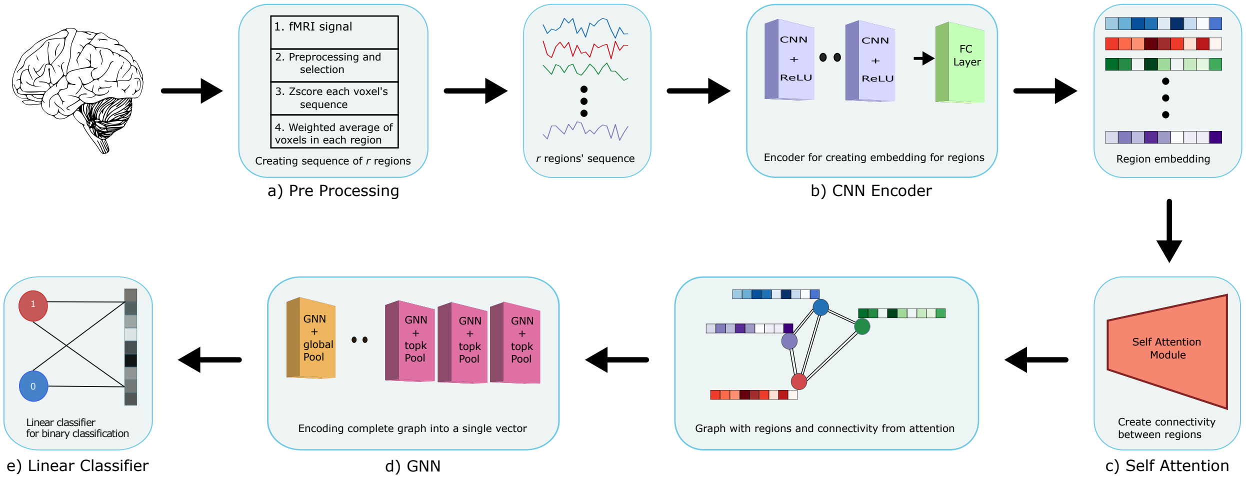

1. fMRI signal
2. Preprocessing and selection
3. Zscore each voxel's sequence
4. Weighted average of voxels in each region
Creating sequence of r regions
a) Pre Processing
r regions' sequence
CNN + ReLU
CNN + ReLU
FC Layer
Encoder for creating embedding for regions
b) CNN Encoder
Region embedding
Self Attention Module
Create connectivity between regions
c) Self Attention
Graph with regions and connectivity from attention
GNN + global Pool
GNN + topk Pool
GNN + topk Pool
GNN + topk Pool
Encoding complete graph into a single vector
d) GNN
1
0
Linear classifier for binary classification
e) Linear Classifier

Figure 4.1 BrainGNN architecture using a) Preprocessing: To preprocess the raw data with different steps (4.1.1). b) 1DCNN: To create embedding for regions (4.1.2.1). c) Self-attention: To create connectivity between regions (4.1.2.2) d) GNN: To obtain a single feature vector for the entire graph (4.1.2.3) and e) Linear classifier: To obtain the final classification.

4.1.2.2 Self Attention

Using the embeddings created by the CNN encoder, the model estimates the connectivity between the regions of the brain using multi-head self-attention following Vaswani et al. (2017) . The self-attention model creates three embeddings namely (key, query, value) for each region, which in the proposed architecture are created using three simple linear layers. Each linear layer ϕ is of size 24. $key_i = \phi_k(region_i)$, $query_i = \phi_q(region_i)$, $value_i = \phi_v(region_i)$. To create weights between a region and every other region, the model takes dot product of a region's query with every other region's key embedding to get scores between them. Hence, $score_{ij} = query_i \cdot key_j$. The scores are then converted to weights using softmax. $w_i = Softmax(score_i)$ where $score_i \in \mathbb{R}^{1\times r}$ is a vector of scores between region i and every other region. The weights are then multiplied with the *value* embedding of each region and summed together to create new representation for a $region_i$. Following equations show how to get new region embedding and weight values.

$$\begin{aligned}
&key_i = region_i * W^{(k)}, \quad value_i = region_i * W^{(v)}, \quad query_i = region_i * W^{(q)} \\
&K = ||_{i=1}^{r} key_i^T = key_i^T||....||key_r^T, \quad weight_i = softmax(query_i * K) \qquad (4.1) \\
&new\_region_i = \sum_{j}^{r}(weight_{ij} * value_j)
\end{aligned}$$

This process is carried out for all the regions, producing new representation of every region and the weights between regions. These weights are then used as the functional connectivity between different regions of brain for every subject. The self attention layer encodes the spatial axis for each subject and provides with the connection between regions. The weights are learned via end to end learning of the model performing classification. This frees from using predefined models or functions to estimate the connectivity.

4.1.2.3 GNN

The graph network used in BrainGNN is based on a previously published model Li et al. (2016). Each subject is represented by a graph G having V, A, E where $V \in \mathbb{R}^{r\times \mathrm{t}}$ is the matrix of vertices, where each vertex is represented by an embedding acquired by self-attention. $A, E \in \mathbb{R}^{r\times r}$ are the adjacency and edge weight matrices. Since the proposed model do not use any existing method of computing edges, a complete directed graph with backward edges is constructed, meaning every pair of vertices is joined by two directed edges with weights e_{ij} and $e_{ji} \in E$. For each GNN layer, at every step s, each node, which is a region sums feature vectors of every other region relative to the weight edge between the nodes and pass the resultant and it's own feature vector through a gated recurrent unit (GRU) network Cho et al. (2014), to obtain new embedding for itself.

$$x_s^{n_i} = \mathrm{GRU}(x_{s-1}^{n_i}, \sum_{\forall n_j : n_j -> n_i} e_{ji} x_{s-1}^{n_j}) \tag{4.2}$$

where GRU can be explained by following set of equations, with h_{s-1} representing the result of sum in Equation 4.2:

$$\begin{aligned} z_s &= \sigma(\mathbf{W}^{(z)} x_{s-1} + \mathbf{U}^{(z)} h_{s-1}) \\ r_s &= \sigma(\mathbf{W}^{(r)} x_{s-1} + \mathbf{U}^{(r)} h_{s-1}) \\ {h_s}' &= \sigma(\mathbf{W} x_{s-1} + r_s \odot \mathbf{U} h_{s-1}) \\ x_s &= \sigma(z_s \odot h_{s-1} + (1 - z_s) \odot h_s') \end{aligned} \tag{4.3}$$

The number of steps is a hyper-parameter and are set as 2 based on the experiments. The graph neural network helps nodes to create new embeddings based on the embeddings of other regions in the graph weighted by the edge weights between them. In the proposed architecture, 6 GNN layers are used, as shown in experiments of Bresson & Laurent (2017) that it provides with the highest accuracy, with the first 3 followed by a top-k pooling layer Gao & Ji (2019); Knyazev et al. (2019). On the input feature vectors which are the embeddings of the regions, the pooling operator learns a parameter ($\mathbf{p}$) which is to assign weight to the features. Based on this parameter, top (k) layers are chosen in each pooling layer and the rest of the regions are discarded from further layers. The pooling method can be explained by the following equations.

$$
\begin{aligned}
y &= \frac{Xp}{\|p\|} \\
i &= \mathrm{top}_k(y) \\
X' &= (X \odot \tanh(y))_i, \\
A' &= A_{i,i}
\end{aligned}
\tag{4.4}
$$

$\mathbf{X}'$ and $\mathbf{A}'$ are the new features and adjacency matrix acquired after selecting top (k) regions. Pooling is performed to help model focus on the important regions/nodes which are responsible for classification. The ratio of nodes to keep in the pooling layer is a hyper-parameter and set as $(0.8, 0.8, 0.3)$ ratios respectively. Since each subject is represented as graph G, in the end graph classification is performed by pooling all the feature vectors of the remaining 23 regions/nodes. To get one feature vector from the entire graph the output of three different pooling layers are concatenated. The complete graph is passed into three separate pooling layers. Each of the pooling layer gives one feature factor. In the end, the three vectors are concatenated to get one final embedding for the entire graph which represents a subject. The proposed architecture uses graph max pool, graph average pool and attention based pool Vinyals et al. (2016). The dimension of the resulting vector is 96. The feature vector is then passed through two linear layers of size 32 and 2. As the name suggests, graph max pool and graph average pool just gets the max and average vector from the graph whereas attention based pooling multiplies each vector with a learned attention value before summing all the vectors.

4.1.2.4 Training and Testing

To train, validate and test the model proposed in this chapter the total 311 subjects are divided into three groups of size 215, 80 and 16, for training, validating and testing respectively. To conduct a fair experiment 19 test folds are created and for each fold 10 randomly-seeded trials are performed, resulting in a total of 190 trials, and selecting 100 subjects per class for each trial. Area under the ROC (receiver operating characteristic) curve (AUC) is calculated for each trial. The model is trained in an end to end fashion, using Cross Entropy to calculate loss by giving true labels Y as targets. Adam is used as the optimizer and reducing learning rate on plateau with patience 10. Early stopping is used with the model based on validation loss, with patience of 15. Let θ represent the parameters of the entire architecture.

$$loss = \text{CrossEntropy}(\hat{Y}, Y) \tag{4.5}$$

$$\theta^{*} = \arg\min_{\theta}(loss; \theta) \tag{4.6}$$

4.2 Results

This section shows three different groups of results. 1) The classification results, 2) Regions' connectivity and 3) Key regions selection. The study discusses these in the following sections. The proposed model is tested and compared against the classical machine learning algorithms and Mahmood et al. (2020b) on the same data used in BrainGNN. The input for the machine learning model is sFC matrices produced using Pearson product-moment correlation coefficients (PCC).

4.2.1 Classification

As mentioned, the AUC metric is used to quantify the classification results of the proposed model. AUC is more informative than simple accuracy for binary classification. Figure 4.2 shows the results for the proposed model. Figure 4.3 shows the ROC curves of the models for each fold. The performance is comparable to state of the art classical machine learning algorithms using hand crafted features and existing deep learning approaches such as Mahmood et al. (2020b), which performed test on independent component analysis (ICA) components with a hold out dataset. Comparison with other machine and deep learning approaches is shown in Figure 4.4 and prove the claim of BrainGNN providing state-of-the-art results. BrainGNN gives almost the same mean AUC as the best performing model i.e. SVM (Support Vector Machine). Presumably, these results are currently among the best on the unmodified FBIRN fMRI dataset Rashid et al. (2016); Saha et al. (2020); Salman et al. (2019). Table 4.1 shows the mean AUC for each cross validation fold that was used for experimentation for BrainGNN. As it is shown in the table that AUC has high variance across the different test sets of cross validation. To make more sense out of the functional connectivity and region selection, both results are based on the second test fold which gives the highest (~ 1) AUC score.

4.2.2 Functional Connectivity

The functional connectivity between regions of the brain is crucial for understanding how different parts of brain are interacting with each other. The proposed model uses the weights assigned by the self-attention module of as the connection between regions. Figure 4.5 shows weight matrices for the second test set in cross validation. Weight matrices of subjects belonging to SZ class turn out to be much sparser than weights of healthy controls subjects. The result shows that the connectivity is limited to fewer regions, and functional connectivity differs across classes and fewer regions get higher weights in case of SZ subjects. Statistical testing is also performed to confirm that the weight matrices of HC differ from those of SZ subjects. For this purpose, two sets each representing the concatenation of the weights of 8 test subjects belonging to a class are created. 2 different testing are performed and shown in Table 4.2. P-value of < 0.0001 shows that we can reject the null-hypothesis, hence making it highly likely that the difference between weights of HC and SZ subjects is not zero. FC matrices produced using PCC method, do not provide such level of information and almost all regions get unit weight between other regions. 4.5 shows the usefulness of learning connectivity between regions in an end-to-end manner while training the model for classification.

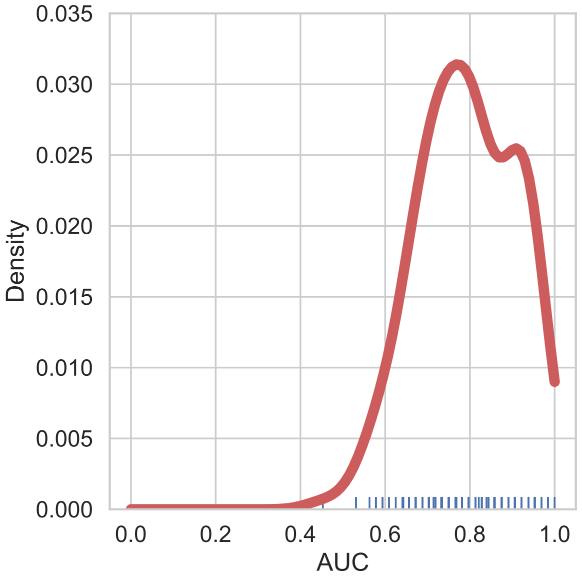


Figure 4.2 KDE plot of probability density of ROC-AUC score on FBIRN dataset. The 190 points on the x-axis signifies the 19 fold cross validation, 10 trials per cross validation. With average and median of ((~ 0.8)), density peaks at (~ 0.8) AUC.

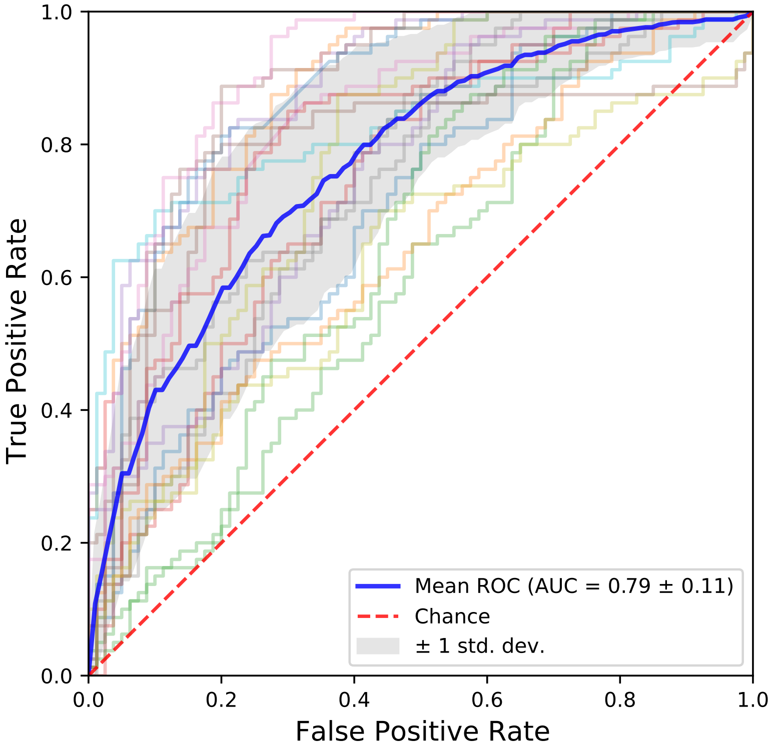


Figure 4.3 Shows the ROC curves of the 19 models generated using each fold of cross validation. The graph is symmetrical and well balanced. It shows that the model did not learn one class over the other.

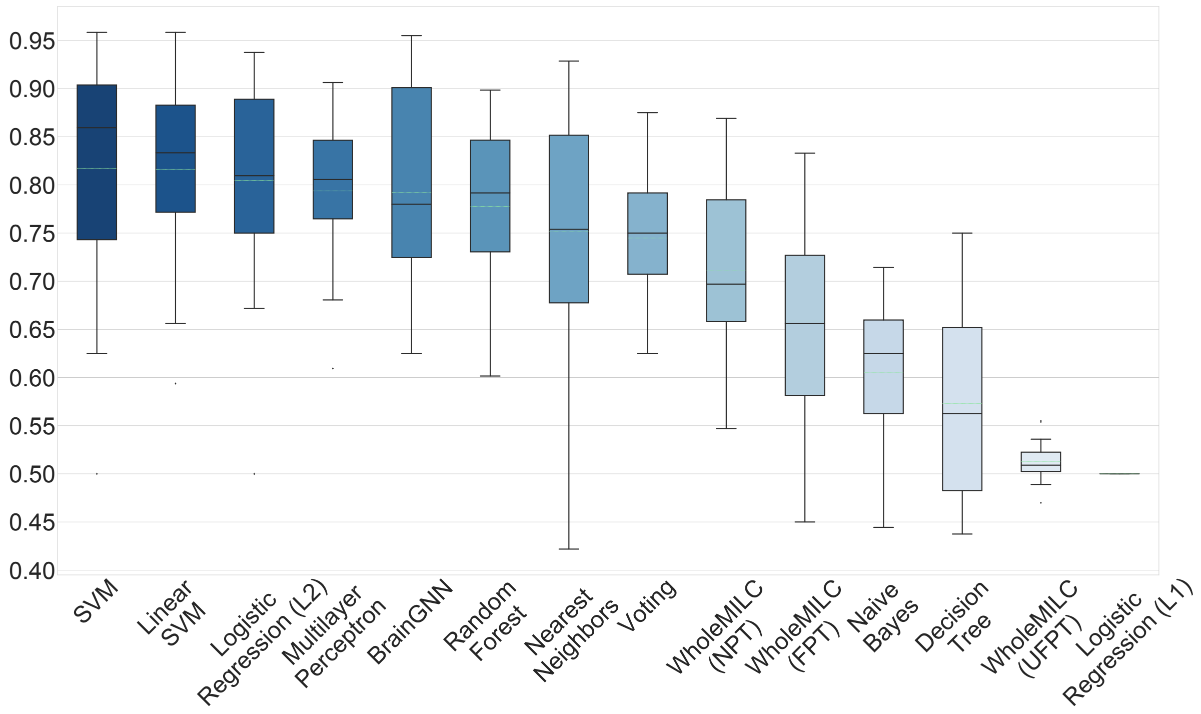


Figure 4.4 BrainGNN comparision with other popular methods. BrainGNN provides mean AUC as 0.79, which is just (~ 0.02) less than the best performing model (SVM). Methods like WholeMILC (UFPT) and l1 logistic regression failed to learn on the input data. The l1 logistic regression model does perform better with a very weak regularization term.

Table 4.1: Showing mean AUC of 10 trials for each cv fold

CV Fold	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
AUC	0.695	0.955	0.644	0.752	0.908	0.917	0.894	0.803	0.649	0.805	0.922	0.699	0.625	0.780	0.794	0.766	0.914	0.750	0.777

4.2.3 Region Selection

The pooling layer added in the GNN module allows to reduce the number of regions. Functionality across brain regions differ significantly and not all regions are affected by a disorder or have any noticeable affect on classification. This makes it very important to know which regions are more significantly informative of the underlying disorder and study how they get affected or affect the disorder. Figure 4.6a shows the final 23 regions selected after the last pooling layer in the GNN model which is just 20 percent of the total brain regions used. The relevance of these regions is further signified by the fact that the graph model has no residual connections and the final feature vector created after the last GNN layer is through the feature vectors of these regions. Figure 4.6b shows the location of the selected regions in the MNI brain space, regions are distinguished by color. Each region is assigned one unit from the color bar, used to represent signal variation in the fMRI data.

Table 4.2: Statistical testing between weight matrices of HC and SZ. The test shows that weights of regions differ across HC and SZ subjects. Refer to Figure 4.4 for mean and deviation of these folds.

Test	P Value
Mann-Whitney U Test	0.0
Welch’s t-test	0.0

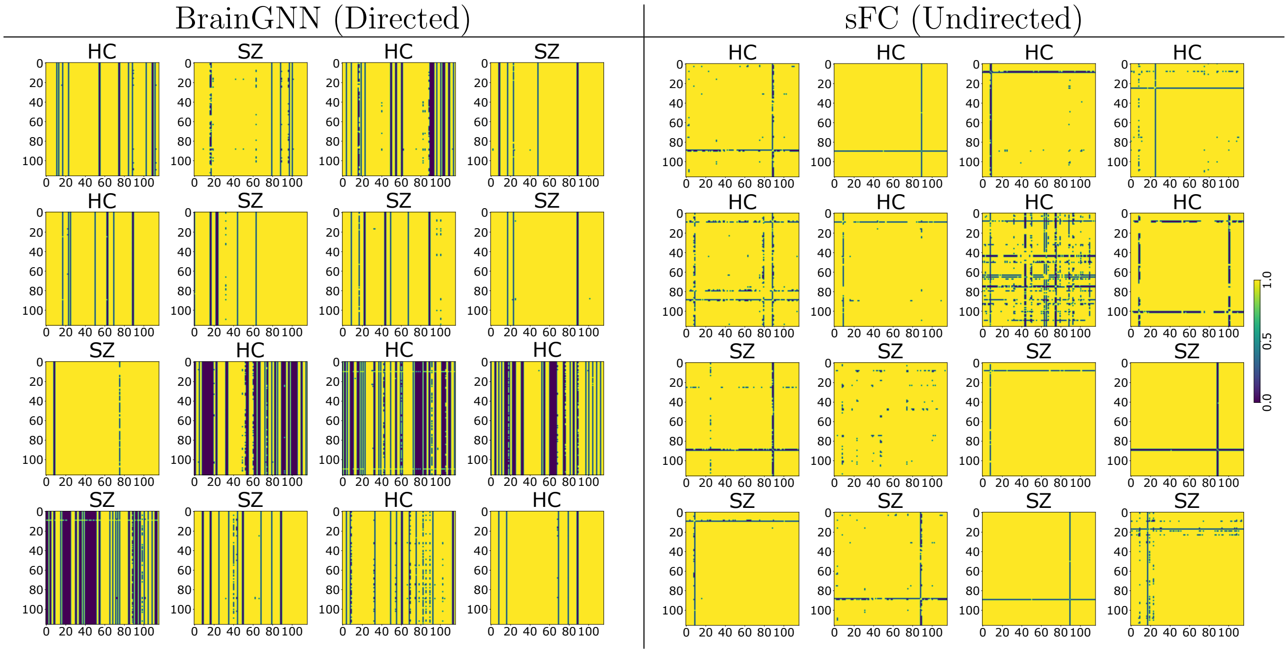
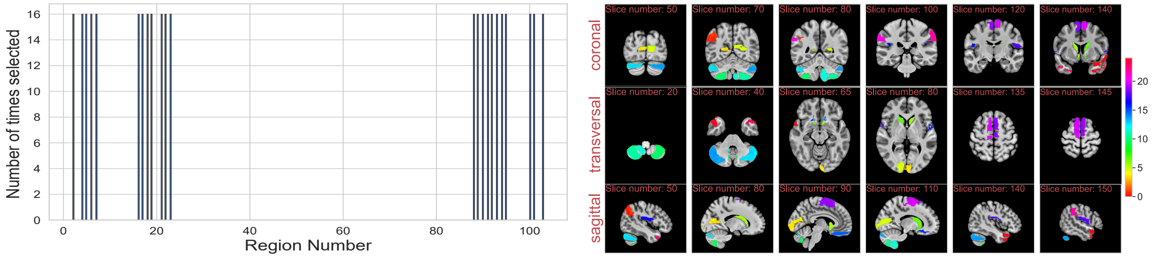


Figure 4.5 Connectivity between regions of subjects of both classes using BrainGNN and sFC (PCC method). **BrainGNN:** The similarity of connection between a class and difference across class is compelling. Weights of SZ class are more sparse than HC, highlighting the fact that fewer regions receive higher weights for subjects with SZ. Refer to Table 4.2 for results of statistical testing between weights of HC and SZ subjects. **sFC**: The matrices are symmetric but are less informative than those produced by BrainGNN. Most of the regions are assigend unit weight.



(a) Region frequency

(b) Region maps

Figure 4.6 4.6a: Histogram of regions selected after the last pooling layer of GNN. 2^{nd} fold of the cross validation gives this figure. All 23 regions are selection equal number of times (16). It further signifies the important of these regions, showing that for all subjects across both classes, these 23 regions are always selection. 4.6b: Mapping the 23 regions back on the brain across the three anatomical planes. 100^{th} time point is selected for these brain scans. X axis shows different slices of the plane.

4.3 Discussion

The richness of results in the three presented categories highlights the benefits of the proposed method. High classification performance shows that the model can accurately classify the subjects and hence can be trusted with the other two interpretative results of the chapter. Functional connectivity between regions shown in the chapter is of paramount importance as it highlights how brain regions are connected to each other and the variation between classes. Learning functional connectivity end-to-end through classification training frees the model from depending on an external method. The sparse weight matrix of subjects with SZ shows that connectivity remains significant between considerably fewer regions than for healthy controls. Notably, the attention based functional connectivity cannot be interpreted as the conventional correlation based symmetric connectivity. Due to the inherent asymmetry in keys and values the obtained graph is directed but is also prediction based rather than simply correlation. It is expected that that a further investigation into the obtained graph structure will bring more results and deeper interpretations. The sparsity is to be further explored and seen in context of the regions selected, shown in the last section of results. The final regions selected by the model strengthens the proposed hypotheses that not all regions are equally important for identifying a particular brain disorder. Reducing the brain regions by almost 80% helps in identifying the important regions for classification of SZ. The regions selected by the proposed model such as (cerebellum, temporal lobe, caudate, SMA) etc have been linked to the disease by multiple previous studies, hence reassuring the correctness of the model Haan et al. (2011); Fu et al. (2021b); Zeng et al. (2017); Culbreth et al. (2021).

We can see an immediate benefit of using GNNs, specifally the proposed BrainGNN model to study functional connectivity. The data-driven model almost eliminates manual decisions transitioning graph construction and region selection into the data-driven realm. With this BrainGNN opens up a new direction to the existing studies of connectivity and it is safe to expect that further model introspection to yield insight into the spatio-temporal biomarkers of brain disorders. The next goal of this dissertation is to make appropriate changes in the proposed model and/or propose new models that can lead to better estimation of the brain network connectivity.

CHAPTER 5
DIRECTED INSTANTANEOUS CONNECTIVITY ESTIMATION

To estimate brain networks' connectivity that is 1) directed, 2) interpretable, 3) flexible, and 4) dynamic, a novel approach is developed and shown in this chapter. The approach is called the Directed Instantaneous Connectivity Estimator (DICE): a predictive model to estimate dynamic directed connectivity between brain networks, represented as a dynamically varying directed graph by predicting the downstream binary label. This proposed model may be placed into the category of model-free connectivity methods as it does not model the data generation process. This chapter defers to using "directed (network) connectivity" (D(N)C) for the graphs that DICE estimates.

Unlike existing supervised DL models that typically produce difficult-to-interpret representations, DICE is designed primarily with interpretability in mind. The proposed model DICE reveals what it learned about the dynamics of brain network connectivity without using post hoc interpretability methods. Effectively, this work has lead to a "glass-box" layer within a traditionally "black-box" DL model. In contrast to commonly used hidden layers, the "glass-box" layer propagates a weighted adjacency matrix of a directed graph, ensuring that it is interpretable in the context of the classification task. Hence, by estimating DC based on the task and using only the estimated connectivity structure for classification, DICE learns to capture task-relevant networks and their connectivity, leading to a flexible estimation of an interpretable DC. By estimating DC instantaneously (window-size = 1), DICE removes the need for the window-size parameter used in many dynamic connectivity studies.

To thoroughly validate DICE's performance, a series of experiments are conducted on four neuroimaging datasets that span three disorders (schizophrenia, autism, and dementia) and cover a wide age range. The model is trained on classification tasks for each of these brain disorders, age prediction, and gender classification, and analyze the resulting DC of the "glass-box" layer. Surprisingly, the deliberate focus on stable interpretable results has an enhancing side effect on DICE's predictive performance. Results show, the model's predictions are better or on par with state-of-the-art methods that were developed with a focus on classification performance rather than interpretability. The model's results show that when learning to classify subjects based on a specific criterion, DICE estimates interpretable DCs specific to that criterion. For gender and mental disorder classification, subgraphs emphasized by the learned DCs are discriminative of gender and mental disorders, respectively. It is also also demonstrated in experiments that DICE learns interpretable DCs distinct to dementia, gender, and age prediction for the same subjects by enhancing connectivity for networks that pertain to the training signal. The flexible estimation of DC structures advances the results of Salehi et al. (Salehi et al. 2020), which show that functional parcel boundaries change for an individual based on the cognitive state. This work shows an increased utility of the inferred directionality for increasing the precision of explainable group differences. As a result, DICE can resolve more states in fMRI dynamics than is resolvable in typical dynamic functional network connectivity analyses. Additionally, DICE incorporates a temporal attention module that highlights crucial time steps relevant to the task, further improving the interpretation of predictions for the dynamics. The learned DC structures and temporal attention weights are stable and consistent across randomly-seeded trials.

5.1 Materials and Methods

5.1.1 Materials

Resting state functional magnetic resonance imaging (rs-fMRI) data as input to DICE. DICE is tested by classifying three different brain disorders, predict gender and age of subjects. For each brain disorder binary classification of healthy controls (HC) and patients is performed. Four datasets used for experiments are collected from FBIRN (Function Biomedical Informatics Research Network [1]) (Keator et al. 2016) project, from release 1.0 of ABIDE (Autism Brain Imaging Data Exchange [2]) (Di Martino et al. 2014) and from release 3.0 of OASIS (Open Access Series of Imaging Studies [3]) (Rubin et al. 1998). Healthy controls from the HCP (Human Connectome Project) (Van Essen et al. 2013) are used for gender prediction. Refer to Table 2.1 for details of the datasets.

5.1.1.1 Preprocessing

Two typical brain parcellation techniques are used to create data for experiments; independent component analysis (ICA) and regions of interest (ROIs) based on a pre-defined atlas. The preprocessing pipeline used depends on the parcellation technique and the pipeline used in state-of-the-art studies for the dataset. Refer to Section 2.1.1 for more details.

[1] FBIRN phase III are used.

[2] `http://fcon_1000.projects.nitrc.org/indi/abide/`

[3] `https://www.oasis-brains.org/`

To get ROIs, two atlases for brain parcellation are used; Shaefer (Schaefer et al. 2017), and Harvard Oxford (HO) (Desikan et al. 2006) with 200, and 111 regions respectively. For each region, average value is computed for all the voxels falling inside a region, thus resulting into a single time-series for each region. After dividing data into regions, each time-series was standardized by their zscore having zero mean and unit variance. All the preprocessing was done before training the model.

5.1.2 Method

DICE receives the time-courses of the ICA components or ROIs represented as a matrix of size $N * T$ (Number of components/ROIs * Number of time-points) and learns a set of T directed graphs representing the dynamic DC or DNC between spatial components (e.g., ICA-based spatial components, regions from an atlas), which are designated as nodes of a graph by predicting the binary labels. Let G represent the set of graphs where $G = \{g_1, g_2, ..., g_T\}$ where T is the total time-points and $g_t = (V_t, E_t)$, where, V_t and E_t represent the nodes and edges present at time-point t. To create the graph g_t, firstly a bidirectional long short-term memory (biLSTM) (Schuster & Paliwal 1997) module is used to create the embedding h_t^i of node i at time t. Then a self-attention module (Vaswani et al. 2017) is used which takes all such embeddings at each time t and create a weight matrix among nodes thus providing the DC (graph) between nodes at each time-point. To create a final graph G^f for downstream classification, a temporal attention model is used that assigns a weight to each g_t and compute the weighted sum of the set G. The working and purpose of each module is explained in detail in the following sections. Figure 5.1 shows the complete architecture.

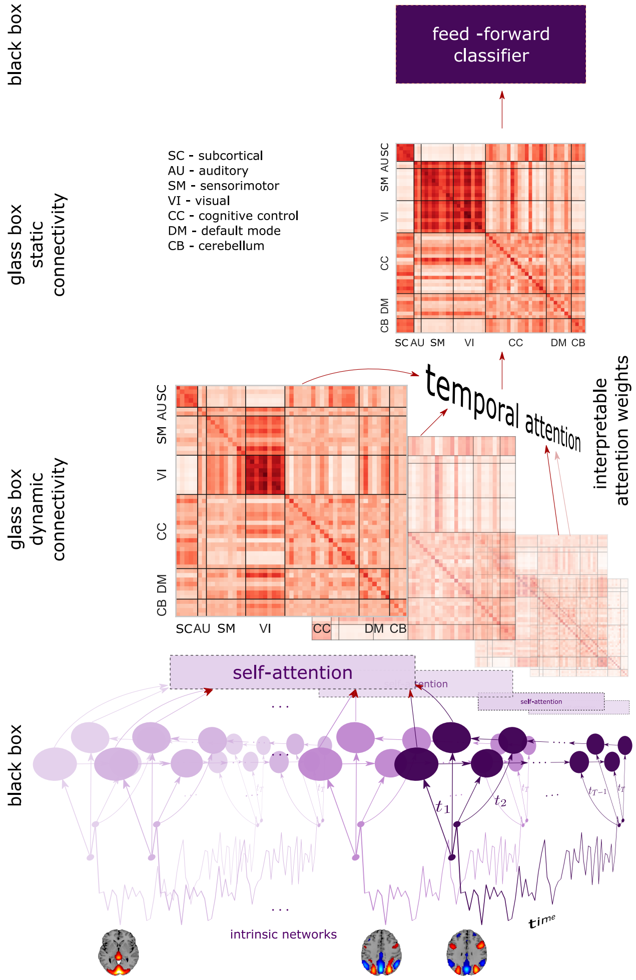


Figure 5.1 DICE architecture using biLSTM, self-attention and temporal attention. The model uses self-attention between the embeddings of all components/nodes at each time-point to estimate the DC $\mathbf{W}_i$. Temporal attention is used to create a weighted sum of the T DC. Architecture details of temporal attention is shown in Figure 5.2.

5.1.2.1 biLSTM

The time-point value x_t^i for node i at time t can be effected by many different factors and relations. Capturing these relations can increase model interpretability and improve downstream classification performance. In a time-series (fMRI data), one of these factors is the values/data at previous time-points $x_{1...t-1}^i$. In fMRI data, this relationship is unknown and is hard to capture and hence cannot be computed using a fixed method/formula (hand-crafted features). The difficulty is further increased by a) low temporal resolution of fMRI data and b) the fact that it is unknown how farther in time the effects of a time-point remains in a time-series. These effects are different for each subject and can even vary among nodes of the same subject. LSTMs have proved to be extremely effective for time-series/sequence data where the model takes an input from a sequence at time-point t and create representation for current and also predict representation for future time-courses based on the representation of previous time-points. LSTMs learn the temporal relationships between data through the cell's memory and forget gate. These gates are optimized on the data and downstream task (ground-truth signal) and the relationships between data are learned instead of computed. The working of the LSTMs can be explained by the following set of equations. σ represents sigmoid activation, b are the biases, and $\odot$ is the Hadamard product (Million 2007).

$$\begin{aligned}
\mathbf{i_t} &= \sigma(\mathbf{W}_{ii}\mathbf{x}_t + b_{ii} + \mathbf{W}_{hi}\mathbf{h}_{t-1} + b_{hi}) \\
\mathbf{f_t} &= \sigma(\mathbf{W}_{if}\mathbf{x}_t + b_{if} + \mathbf{W}_{hf}\mathbf{h}_{t-1} + b_{hf}) \\
\mathbf{g_t} &= \tanh(\mathbf{W}_{ig}\mathbf{x}_t + b_{ig} + \mathbf{W}_{hg}\mathbf{h}_{t-1} + b_{hg}) \\
\mathbf{o_t} &= \sigma(\mathbf{W}_{io}\mathbf{x}_t + b_{io} + \mathbf{W}_{ho}\mathbf{h}_{t-1} + b_{ho}) \\
\mathbf{c_t} &= \mathbf{f_t} \odot \mathbf{c_{t-1}} + \mathbf{i_t} \odot \mathbf{g_t} \\
\mathbf{h}_t &= \mathbf{o_t} \odot \tanh(\mathbf{c_t})
\end{aligned} \tag{5.1}$$

Here $\mathbf{h}_t$ is the representation/embedding for the input at t. The model uses a biLSTM to create representation $\mathbf{h}_t$ for each node i. Thus $\mathbf{h}_t^f = LSTM(x_t, \mathbf{h}_{t-1})$, $\mathbf{h}_t^b = LSTM(x_t, \mathbf{h}_{t+1})$ and $\mathbf{h}_t = \text{concatenate}(\mathbf{h}_t^f, \mathbf{h}_t^b)$. Here $\mathbf{h}_t^f$ and $\mathbf{h}_t^b$ are representation for forward and backward pass. LSTM is used for each node (component/region) individually, sharing weights of LSTM among the nodes. x_t^i (scalar value) is given as input to the LSTM along with hidden vector and receive $\mathbf{h}_t^i$ for the node i, which solves the window size problem occurring in dynamic-FNC studies. To make it easier to understand, one can assume that in DICE the window size is 1. This allows the model to later instantaneously compute connectivity matrix (links/edges) between the nodes at each time-point. The biLSTM receives temporal values of each component/region separately but share the weight matrices across regions. This allows the biLSTM to learn the temporal connections by looking at multiple nodes but does not learn spatial dependencies among nodes. For this exact reason self-attention across nodes is used in DICE.

5.1.2.2 Self-Attention

A node in a graph can be linked with other nodes represented as the edge connectivity between them. The connectivity between nodes influence the value of a node (x_t^i) at a certain time-point. Thus it is important to measure the connectivity between nodes for the construction and interpretation of the graph. In fMRI data where each x^i is a brain region/component, capturing the DC or DNC between nodes shows how brain networks are linked with each other and the direction of flow of information between brain networks. The estimated matrices can then be used to explain brain working and brain disorders. Connectivity between brain regions is independent of the structural connectivity and thus is unknown. To capture the directed connectivity between brain regions, a self-attention module is used in this work.

Self-attention module captures the weights between n inputs of a sequence. Since in a dynamic system (brain network), the connectivity between nodes can change at any instance, therefore, at each time-point t a sequence of n vectors $\mathbf{h}_t^1 ... \mathbf{h}_t^n$, $n =$ total nodes, is passed as input to the self-attention module and create the weight matrix $\mathbf{W}_t$, where each $\mathbf{W}_t \in \mathbb{R}^{n*n}$ is the connectivity weight matrix of input nodes at time-point t.

The self-attention module creates three embeddings, namely, key ($\mathbf{k}$), value ($\mathbf{v}$), and query ($\mathbf{q}$) and creates new embeddings for each input using these embeddings. The following set of equations can sum up the whole process. For simplicity, the t is omitted from these equations. $^\intercal$ represents transpose and $\oplus$ represents concatenation.

$$\begin{aligned}
&\mathbf{k}^i = \mathbf{h}^{i\intercal}\mathbf{W}^{(k)}, \quad \mathbf{v}^i = \mathbf{h}^{i\intercal}\mathbf{W}^{(v)}, \quad \mathbf{q}^i = \mathbf{h}^{i\intercal}\mathbf{W}^{(q)} \\
&\mathbf{K} = \oplus_{i=1}^{n}\mathbf{k}^{i\intercal}, \quad \mathbf{w}^i = \mathrm{softmax}(\mathbf{q}^i\mathbf{K}) \\
&\mathbf{W} = \oplus_{i=1}^{n}\mathbf{w}^i
\end{aligned} \tag{5.2}$$

Here $\mathbf{W} \in \mathbb{R}^{n*n}$ is the connectivity matrix between n nodes in the graph. As brain disorder are associated with disruptions in the connectivity of brain's intrinsic network, only the learned directed connectivity matrices $\mathbf{W}$ are used for downstream classification and not the features, thus forcing the model to estimate the differences in connectivity between the two classification groups (e.g., HC and patients). As DICE is tuned to estimate the DC or DNC for the groups of subjects and output the it, DICE captures and shows the basis of downstream classification. The DC or DNC estimated by the model can be easily represented as a graph which are extremely easy to interpret. The self-attention glass-box layer shows task-dependant nodes (brain regions) and their connectivity.

The features that represent time-courses are used to learn/estimate the DC or DNC structure. As the true connectivity/graph structure is never available in many applications to directly compare with, this study proposes that a connectivity matrix leading to state-of-the-art classification performance makes it more reliable than using the representa-

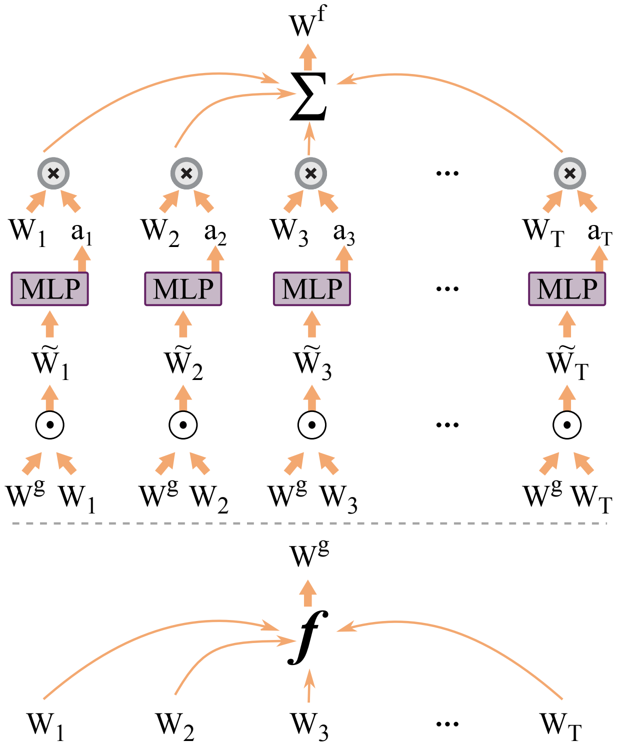


Figure 5.2 GTA architecture for temporal attention. $\mathbf{W}_{1-T}$ matrices are summed to create $\mathbf{W}^{global}$. Using $\mathbf{W}^{global}$ and $\mathbf{W}_i$ attention score α_i is created for each time-point. Refer to equations in 5.3 and 5.4 for working details. Here $\boldsymbol{f}$ denotes the average function.

tions/embeddings for classification.

5.1.2.3 Temporal Attention

As the model uses only the learned connectivity matrices for downstream classification. For this purpose, there is a need to create a single weight matrix W^f based on the W_{1-T} matrices. For the downstream classification task, not all the time-points are equally important, hence it is crucial to incorporate a temporal attention module which assigns weight to each $\mathbf{W}_t$ and calculate a weighted average of all the weight matrices. This section introduces a novel temporal attention module which is called global temporal attention (GTA). **GTA:** To give the attention module a global view of the graph, this work presents a new method called GTA. The global view allows the model to learn how each DC contributes to the global graph or structure of the data in the downstream task. GTA module creates an average of all the T DC and call it $\mathbf{W}^{global}$ representing the global view. Then the similarity of each local $\mathbf{W}_t$ with the global view is compared and used them to create the temporal attention vector $\boldsymbol{\alpha}$. Figure 5.2 shows the architecture details.

$$\begin{aligned}
&\mathbf{W}^{global} = \tfrac{1}{T} \textstyle\sum_{t=1}^{T} \mathbf{W}_t \\
&\widetilde{\mathbf{W}}_t = \mathbf{W}_t \odot \mathbf{W}^{global} \\
&\boldsymbol{\alpha} = (\oplus_{t=1}^{T}((((\mathrm{flat}(\widetilde{\mathbf{W}}_t))\mathbf{W}^{MLP_{l1}})\mathbf{W}^{MLP_{l2}})\mathbf{W}^{MLP_{l3}})
\end{aligned} \tag{5.3}$$

Here $\odot$ is the Hadamard product (Million 2007) between matrices. $\mathbf{W}^f$ is computed as:

$$\mathbf{W}^f = \sum_{t=1}^{T} \mathbf{W}_t \alpha_t \tag{5.4}$$

5.1.3 Training

GTX 2080 with PyTorch as ML framework is used for the experiments. The hidden dimensions for the biLSTM was set to 100, whereas, self-attention including key, query, and value modules, were all set to 48. The dimensions of multi-layer perceptron (MLP) layers for calculating temporal attention vector were $\eta_1 * len(flat(\mathbf{W}_t))$, $\eta_2 * len(flat(\mathbf{W}_t))$, and 1 with $\eta_1 = \eta_2 = 0.05$. It was noticed in the experiments that multiple heads of self-attention increases stability of the estimated DC. Batch normalization is used after the first MLP layer. ReLU activation was used in DICE between the MLP layers. A final two-layer MLP was used to get logits for binary classification problem with $\mathbf{W}^f$ as input with dimensions 64 and 2. Cross-entropy loss with Adam optimizer was used to calculate loss and optimize the model during training. Let θ represent the parameters of the entire architecture, $\hat{y}$ being the predictions and y the true labels, the loss is calculated as:

$$loss = \text{CrossEntropy}(\hat{\mathbf{y}}, \mathbf{y}) + \lambda \|\boldsymbol{\theta}\|_1 \tag{5.5}$$

$$\boldsymbol{\theta}^{*} = \arg\min_{\boldsymbol{\theta}}(loss) \tag{5.6}$$

Additional experiments were also done with an additional loss terms to encourage the model to estimate connectivity matrices where the values of the main diagonal are closer to 1. Please refer to Section B for details. L1-regularization was used to get a sparser solution. λ (regularization weight) was set as $1e^{-6}$ and learning rate was $2e^{-4}$. Based on the experiment, learning rate was reduced either when validation loss reached plateau by a factor of 0.5 or exponentially with $\gamma = 0.99$. Early stopping was used to stop training the model based on validation loss and patience of 25. For each dataset (ICA components or ROIs), to have a fair result, n-fold testing was performed where the value of n depended on the dataset and methods the model was compared against. For each test fold experiments were performed with 10 randomly-seeded trials. This study reports the mean AUC-ROC (Area Under Curve - Receiver Operating Characteristic) across the n test folds and the 10 randomly-seeded trials as it is a more reliable metric than simple accuracy for binary classification tasks. For example, for FBIRN data there were 18 test folds and for each fold 10 trials were performed, which gave a list of 180 AUC-ROC values and the average of these values are reported to show classification performance of the model. In some cases metrics as well, such as accuracy are also reported. Due to the size of the data, there was a need to make some hyper-parameter changes for HCP region-based (ROIs) experiments. The hidden dimension size for bilstm and self-attention module was set to 64 and 32. η_1 was set to 0.005. Furthermore, because of memory constraints encountered during HCP region experiments, during both training and testing total time-points (1200) were divided into a set of three, each having 400 time-points. Logits were acquired for all and the mean was computed to get final logits. Batch size was set to 32.

5.1.3.1 Hyper-parameters Selection and Fine-tuning

All the parameters (hidden dimensions, number of layers, η_1, η_2, λ, learning rate, γ, patience, batch size) mentioned in section 5.1.3 were set as hyper-parameters. The hyper-parameters were fin-tuned based on the average performance of the model on validation dataset across all the folds. Hyper-parameters tuning was not based on the test folds and only test-set results are reported. Notably, for the experiments on DICE, the order of subjects for each dataset was permuted and the experiments were performed using the permuted order. This was done to avoid imbalance of subjects in the folds. On the same lines, when dividing the data into n-folds (test folds) the number of subjects of both classes in each fold were balanced. For example, in case of FBIRN data with 311 subjects and 151 and 160 subjects in class 0 and 1 respectively. When performing 18 fold testing, each test fold consisted of $\lfloor \frac{151}{18} \rceil$ subjects from class 0 and $\lfloor \frac{160}{18} \rceil$ subjects from class 1 and the rest of the data was used for training and validation, where the validation set size was kept same as the test set size. The validation set was used for hyper-parameters tuning, early stopping during training and selecting the model to apply on the test data. It was made sure that no subject (or sessions of a subject) repeated across training, validation and test sets. The exact size of training, validation and test set can be calculated using the criteria mentioned above and the total number of subjects and number of folds mentioned in Table 2.1. In some of the experiments keeping the same number of subjects in each fold created a small data leakage at the end. For the results reported, the maximum leakage was for FBIRN dataset with 18 test folds. For this purpose, another experiment was performed on FBIRN dataset where the last fold had all the left out subjects to prevent any data leakage. This had no effect on the performance of the model. Refer to Table A.1 for results with different folds.

5.2 Experiments

To test if DICE accomplishes all the goals, this study performs detailed experiments by classifying three brain disorders, classify male and female groups for HCP and OASIS subjects, and predict age for OASIS subjects. Experiments for all datasets were performed using ICA time-courses and experiments on FBIRN, ABIDE and HCP data were performed using regions-based (ROIs) data as well. Average results for all the trials are reported. Depending on the experiment, the classification results are compared with state-of-the-art DL methods (Mahmood et al. 2019a, 2020a, 2021b; Gadgil et al. 2021; Kim & Ye 2020; Zhang et al. 2018a; Weis et al. 2019; Arslan et al. 2018) and ML methods (Support Vector Machine (SVM), Logistic Regression (LR)). To avoid any discrepancy the results of the DL methods are taken directly from the published studies, even though some studies use test data instead of validation data for selecting the best performing model/parameters. For ML methods the python package Polyssifier[4] was used which selects the best model/parameters based on the performance on validation data.

To show the efficacy of DICE, the acquired results are divided into three broad categories. The following sections show a) classification performance of DICE, b) learned DC and DNC and c) the effects of temporal attention module.

5.2.1 Classification

Figure 5.3 shows the classification performance of DICE using ICA data, Table 5.1 shows the performance using region-based (ROIs) data of FBIRN and HCP, and Table 5.2 shows results on ABIDE region-based (ROIs) data.

[4] `https://github.com/alvarouc/polyssifier`

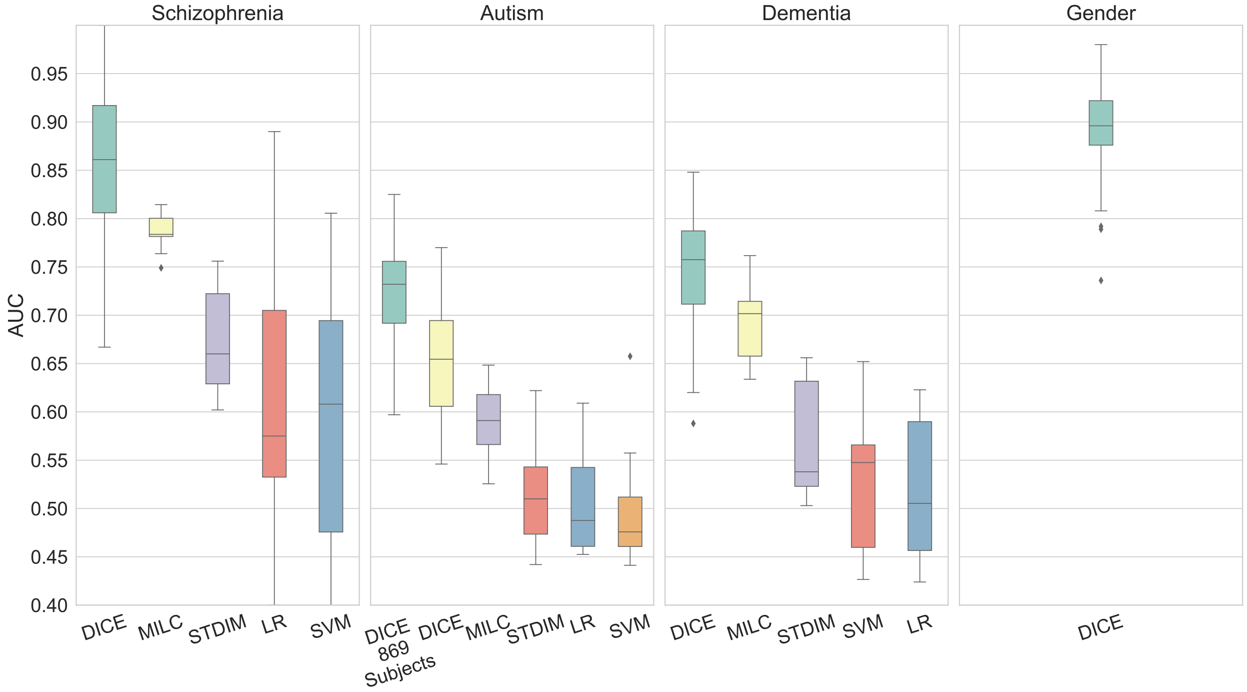


Figure 5.3 AUC comparision of DICE model with four different methods (MILC (Mahmood et al. 2020a), STDIM (Mahmood et al. 2019a), logistic regression (LR), support vector machine (SVM)), over four different datasets using ICA time-courses (Ref to section 2.1.1.1). The proposed method significantly outperforms SOTA methods. Autism experiments with 869 subjects (all TRs) were performed as well. As DICE does not have a pre-training step, NPT is compared with not-pre-trained (NPT) version of MILC and STDIM. Input to ML methods were the same ICA time-courses, not the FNC matrices. Any notable study for gender classification of HCP subjects using was not found that used ICA components as notable methods used ROIs based data. The results using ROIs are shown in Table 5.1. 18 test folds were used for Schizophrenia experiments, all other experiments were done with 10 test folds. Refer to Section A for results on different number of test-folds.

Table 5.1: Classification performance comparison of DICE with other DL methods on region-based (ROIs) data of HCP and FBIRN datasets (Ref to section 2.1.1.2). Our DICE model outperforms all other methods in almost every metric. The best two scores are shown as bold and italic respectively. **Note:** As DICE uses all the regions in the atlas the mean accuracy for SVM-RBF Weis et al. (2019) is reported in this table. The results for GCN Arslan et al. (2018) on HCP data are reported by GIN Kim & Ye (2020). GIN Kim & Ye (2020) and ST-GCN Gadgil et al. (2021) use test data as validation data for choosing the best performing model. It is worth noting that a newer version of GIN Kim & Ye (2020), named STAGIN Kim et al. (2021) reports AUC and ACC score of 92.96 and 88.20 respectively using 1093 subjects, and 5-fold testing. STAGIN Kim et al. (2021) reports much lower ACC for GIN and ST-GCN (81.34 and 76.95 respectively) when not using test data as validation data and keeping other parameters (data, preprocessing, parcellation etc.) same. NA: Not Available.

	HCP - Gender Classification						FBIRN	
	DICE	GIN	SVM-RBF	GCN	ST-GCN	PLS	DICE	BrainGNN
AUC	**0.935**	NA	NA	NA	NA	*0.881*	**0.825**	*0.788*
ACC (%)	**85.8**	*84.6*	68.7	83.98	83.7	79.9	NA	NA
Precision (%)	*85.7*	**86.19**	NA	*84.59*	NA	NA	NA	NA
Recall (%)	**90.2**	86.81	NA	*87.78*	NA	NA	NA	NA
Parcellation	Shaefer 200	Shaefer 400	Shaefer 400 + Fan 39	Shaefer 400	Multi-modal 22	Dosenbach 160	Shaefer 200	AAL 116
Test Folds	10	10	10	10	5	10	18	18
Subjects	942	942	434	942	1091	820	311	311
Study	Our	Kim & Ye (2020)	Weis et al. (2019)	Arslan et al. (2018)	Gadgil et al. (2021)	Zhang et al. (2018a)	Our	Mahmood et al. (2021b)

Table 5.2: Comparison of AUC score on ABIDE region-based (ROIs) dataset. Existing methods use Harvard Oxford (HO) parcellation with 111 brain regions, therefore DICE was tested DICE using two atlases. Unlike Parisot et al. (2018); Cao et al. (2021) DICE uses only fMRI data. We can see that DICE model doesn't depend on the region atlas and gives similar performance using different atlases for region parcellation of the brain. 10 test folds were used for DICE experiments.

Method	Parcellation	Input	n_regions	AUC
DICE	Shaefer	fMRI data	200	*0.70*
DICE	HO	fMRI data	111	0.69
GCN Parisot et al. (2018)	HO	fMRI + phenotypic data	111	**0.75**
DeepGCN Cao et al. (2021)	HO	fMRI + phenotypic data	111	**0.75**
Metric Learning Ktena et al. (2018)	HO	fMRI data	111	0.58

DICE beats every state-of-the-art method used for comparison in this work in almost every metric for both ICA and region-based (ROIs) fMRI data across all datasets when using similar input data (fMRI). As DICE does not use phenotypic information about subjects, it lacks behind (Parisot et al. 2018; Cao et al. 2021) on ABIDE. Parisot et al. (Parisot et al.

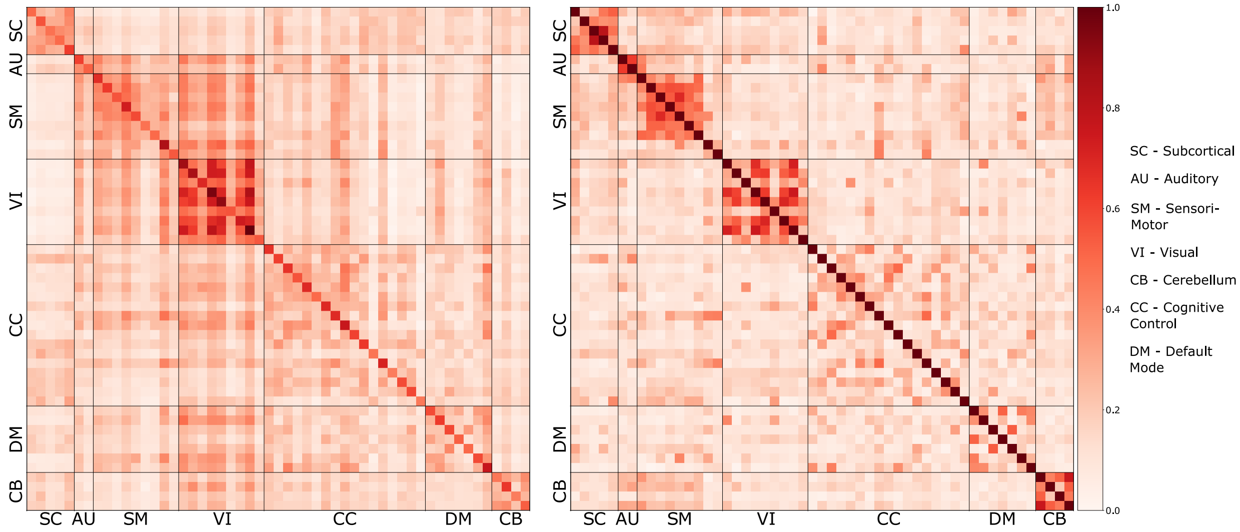
2018) reports a decrease of ~ 2.5 AUC by using a different phenotypic information which clearly shows the dependence on phenotypic data. Whereas, Ktena et al. (Ktena et al. 2018) reports much lower AUC score by using only fMRI data. ML methods fail completely even on ICA data, this study attributes this failure to two reasons. 1) The number of dimensions (m) being much higher than the number of subjects (n), thus creating the curse of dimensionality ($m >> n$) and 2) The ML methods do not compute a graph structure for estimating the connectivity between the networks/components and instead mostly work with independent networks/components. Presumably, no other model gives such high classification score across four neuroimaging datasets. The high classification score of the model computed using only the learned DC structure increases the confidence in the correctness of the learned DC structures.

5.2.2 Directed Connectivity

The learned interpretable, task-dependent (flexible) directed connectivity structures by DICE is the most important contribution of this work. As this is a novel work, this chapter shows in detail, different aspects of the learned connectivity structures. This section a) compares the learned DNC with FNC computed via PCC, b) compares the differences in DC and DNC between multiple classification groups, c) shows how direction matters in connectivity, something which is not captured by FC and FNC, d) dives into the fact mentioned in introduction that unlike computed FNC (using PCC) the learned DNC of DICE is task dependent and changes based on the downstream task (ground-truth signal) and e) shows the dynamic connectivity states for FBIRN data for HC and schizophrenia (SZ) subjects. All the aspects (a-e) discussed in detail in following sections show the correctness and interpretability of the learned DC and DNC. The interpretability of the connectivity matrices estimated by DICE give insight into how brain networks are linked with each other and with the downstream classification task. This is very crucial to understand brain disorders and relevant brain networks. Unlike typical FC and FNC which ranges from -1 to 1, DICE's learned matrices are based on attention and hence ranges from 0 to 1. More information on this in appendix B.

5.2.2.1 DNC vs FNC

As the true connectivity between brain networks is not known, the learned DNC is compared with FNC. Figure 5.4 shows the DNC learned by DICE and the FNC computed using PCC using ICA components for FBIRN dataset. The DNC is $\mathbf{W}^f$ explained in section 5.1.2.3. Both DNC and FNC is the mean matrix for highest performing fold of FBIRN dataset with 16 subjects. The 100 ICA components are divided into informative (53) and noise (47). This section shows the connectvity between 53 non-noise components. These components are further divided into 7 domains/networks following (Allen et al. 2011a). Both matrices clearly show high intra-domain connectivity. The learned DNC shows similar pattern of FNC which increases the confidence in the DNC learned by DICE but there are very important differences between the two. **Inter-network connectivity:** We can see that the estimated DNC finds much more inter-network connectivities than the FNC which is mostly intra-network and has very low scores between networks. **Directionality:** Regarding the direct influence, DNC estimated by DICE is directed and shows components in visual affecting components through out the domains, such information is not present in the FNC which is un-directed (symmetric across main diagonal) and does not show the direction of connectivity. Refer to section 5.2.2.2 for more detail on this.



(a) DICE DNC

(b) PCC FNC

Figure 5.4 Here comparison is done between the estimated DNC with computed FNC using PCC method. 5.4a is the connectivity matrix generated by DICE for FBIRN dataset. A test fold of 16 subjects was used and computed mean DNC for all subjects (10 trials per subject). 5.4b is the mean connectivity matrix of the same subjects generated by PCC. Both figures show similar intra-network connectivity patterns, which verifies the correctness of the connectivity matrix learned by DICE. DICE's estimated DC is directed and captures more inter-network connectivity than FNC. To match the positive weights of DICE, the FNC matrices are normalized from 0 to 1 instead of -1 to 1.

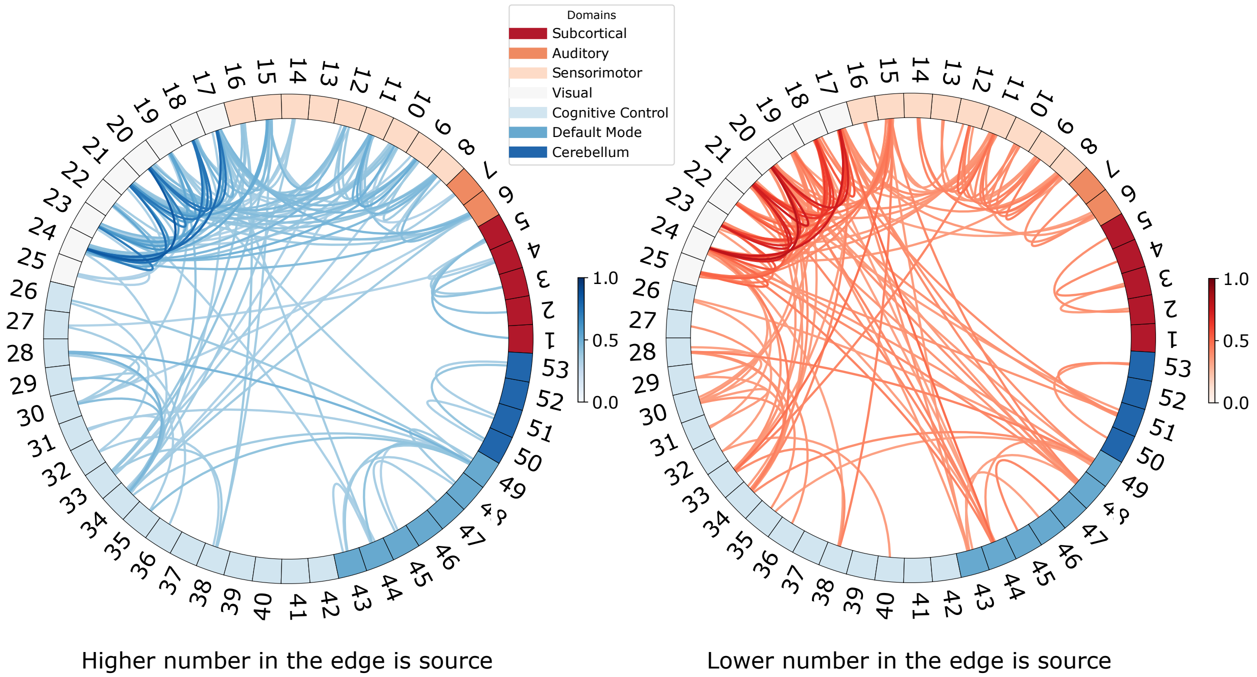
To compare the connectivity matrices in terms of classification results, this study uses an LR model by giving PCC-based FNC and the learned DNC as input. Refer to Table 5.3 for comparison.

Table 5.3: D/FNCs matrices are compared on the basis of AUC score on FBIRN dataset. A logistic regression (LR) model is trained using FNCs computed by PCC, and using DNCs estimated by DICE. Performance using estimated DNCs is in reaching distance of ML methods using hand-crafted features (FCs). A show some experiment details that lead to an even improved classification results.

Method	Input	Mean	Max	Min	Std Dev
LR	PCC FNC	0.883	1	0.72	0.085
LR	Our DNC	0.86	1	0.62	0.096

5.2.2.2 Directed Connectome

Capturing directed connectivity is one of the methods to understand the direction and flow of information in the brain. Learning the direction of connectivity is one of the main advantages of DICE as it might explain the direct influence of brain networks upon each other. To show the direction between components, the estimated DNC of FBIRN subjects is divided into two connectomes showing the direction. Figure 5.5 left shows the edges from a to b where $a > b$. For example the edge between $(8, 23)$ shows the edge from 23 to 8, whereas, Figure 5.5 right shows the opposite. It is clear from the figure that direction matters and the connectivity between brain regions is beyond simple statistical dependence. For example, Figure 5.5 shows that the components in visual network (VIN) affect components in other networks and the edges in the opposite direction are relatively much fewer. We also see direction of connectivity from cognitive control (CC) to sensorimotor (SM). Existing studies (Breukelaar et al. 2017; Cole & Schneider 2007; Tsai et al. 2019) show that cognitive control is responsible for activities like attention, remembering and execution, things which are required when doing a motor task controlled by sensorimotor. Such directionality is important to study brain's working in more detail and is not present in FNC used by existing methods. The results are further discussed in section 5.3.1



Higher number in the edge is source

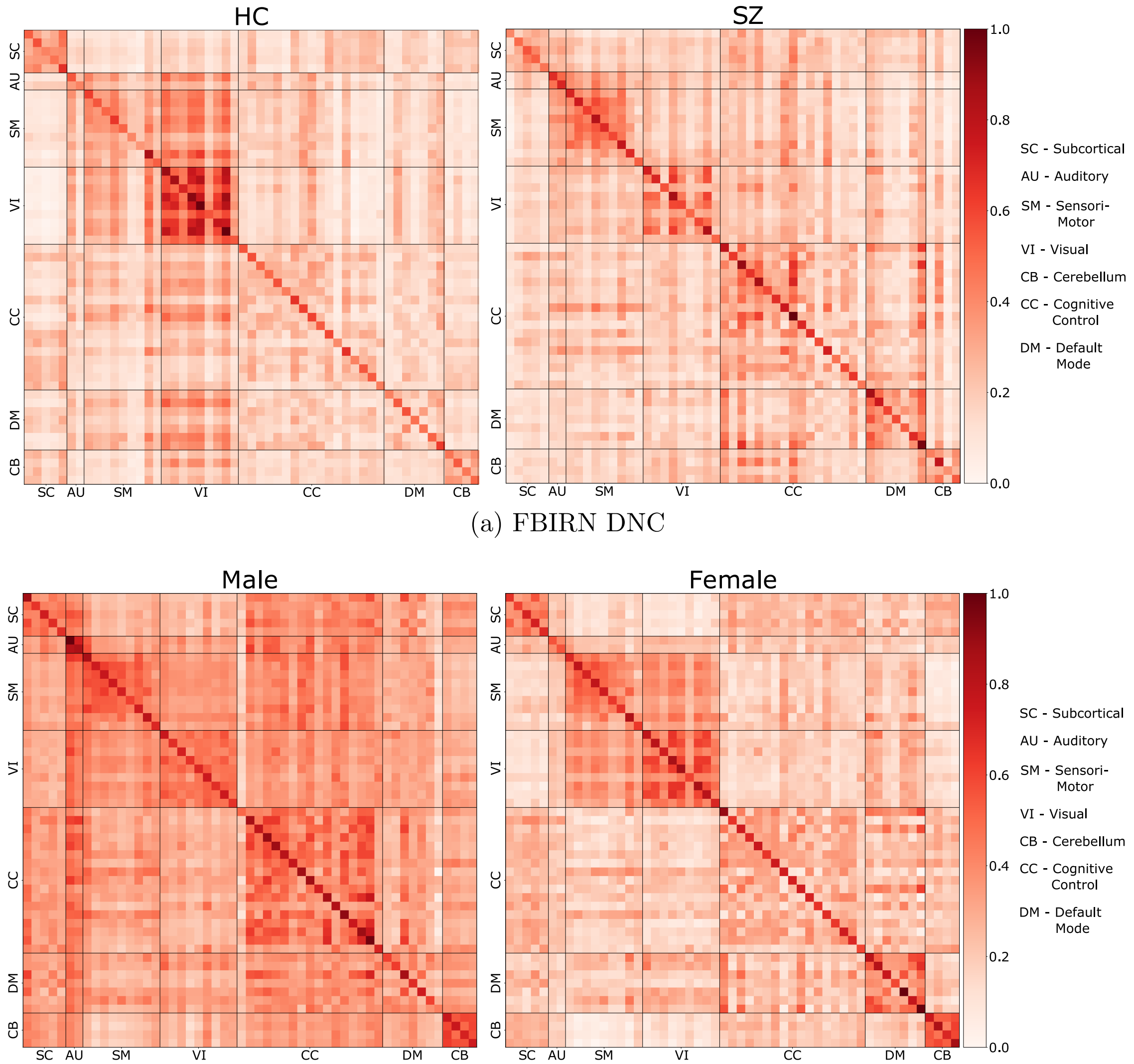
Lower number in the edge is source

Figure 5.5 The figure shows the top 10% directed edges of FBIRN DNC. The numbers represent the 53 non-artifact components. The figure clearly shows the high intra-domain connectivity which matches the existing literature. Direction clearly matters as visual components affect other components but not the opposite way. The direction of edges between CC and SM networks is also of significance. **Edges:** VI → other: 79, other → VI: 25. CC → SM: 9, SM → CC: 3.

5.2.2.3 Connectivity Differences Among Groups

As hypothesized that brain disorders are linked with the connectivity of brain's intrinsic networks, this section shows how the learned DC and DNC changes for subjects belonging to different groups. Figure 5.6a shows the DNC estimated by DICE of HC and SZ subjects for FBIRN data whereas Figure 5.6b shows DNC of male and female groups for OASIS dataset. Both results are computed using ICA pre-processed data. For ICA based DNC, there are similarity between the two matrices as they come from the same joint ICA. However, there are visible difference between the two for multiple networks like visual (VI), cognitive control

(CC), default-mode (DM) and cerebellum (CB). The biggest difference between HC and SZ groups seems to be in the connectivity strength for VIN. For OASIS results 5.6b we can see that females show high connectivity scores in default-mode network (DMN) compare to males and low sensori-motor network (SMN) connectivity compare to males, this has been verified by existing studies (Kim et al. 2021; Filippi et al. 2013; Mak et al. 2016; Ritchie et al. 2018). To verify this by numbers, this study uses statistical testing to compare the two groups (male, female) and compare average connectivity for male and female in DMN and SMN. Table 5.4 shows the statistical results.



(a) FBIRN DNC

(b) OASIS DNC

Figure 5.6 DNC matrices across the binary classification groups using ICA data are compared in this figure. Figure 5.6a is the estimated DNC on FBIRN data for HC and SZ patients. We can see high inter and intra-connectivity in SM and VI networks for HC, which is missing in SZ patients. Figure 5.6b compares DNC between male and female groups using OASIS data. Female group shows hyper-connectivity in DMN and hypo-connectivity in SMN when comparing to male groups.

Table 5.4: Shows stats between male and female DNCs (5.6b) estimated using ICA time-courses of OASIS . We can see that the estimated DNCs for male and female subjects are highly significantly different. For females DMN is hyper-connected than SMN whereas for male SMN has higher average connectivity score than DMN. This shows that the model accurately captures the group differences among male and female subjects and uses the connectivity difference in DMN and SMN to classify male and female subjects. F - Female, M - Male, All - All networks/complete matrix. Results of classification performance is shown in Table 5.8. Table 5.5 shows the p-value significance ranges.

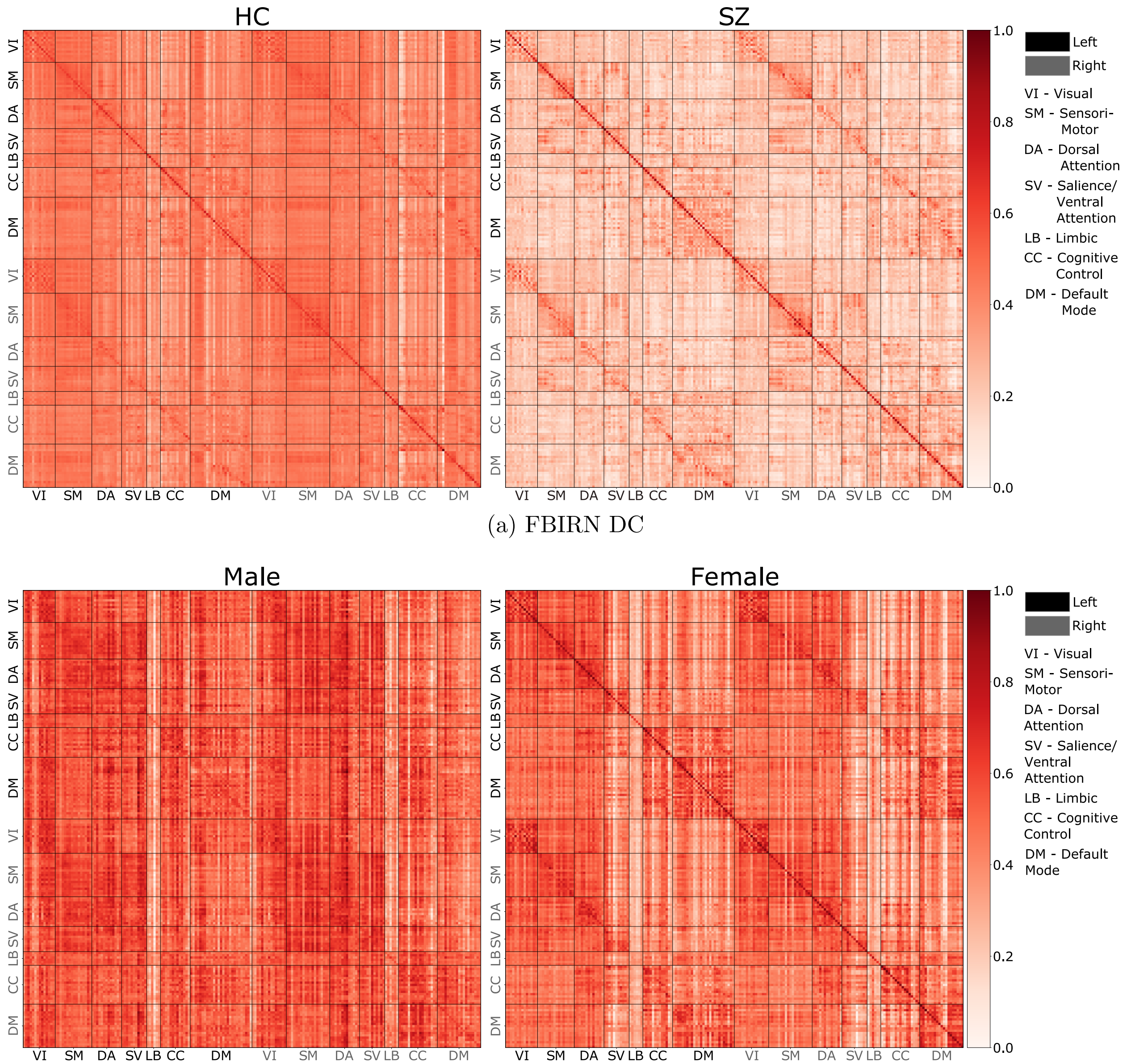
Network 1	Network 2	Test Type	P-value	Avg. Connectivity 1	Avg. Connectivity 2
F_All	M_All	t-test	1e-250	0.353	0.311
		manwhitneyu	1e-256		
F_DM	F_SM	t-test	0.15	0.536	0.510
		manwhitneyu	0.12		
M_DM	M_SM	t-test	5e-5	0.417	0.575
		manwhitneyu	4e-5		
F_DM	M_DM	t-test	6e-4	0.536	0.417
		manwhitneyu	4e-4		
F_SM	M_SM	t-test	3e-4	0.510	0.575
		manwhitneyu	5e-5		

Table 5.5: Ranges of p-value and the corresponding significance level. ns (no significance).

P-value	$p > 0.10$	$0.05 < p < 0.10$	$0.01 < p < 0.05$	$0.005 < p < 0.01$	$0.0001 < p < 0.005$	$p < 0.0001$
Significance	ns	*	**	***	****	*****

Figure 5.7 performs the same experiment for region-based (ROIs) data. Here the regions for both sides of the brain (left and right) are divided into 7 domains following shaefer (Schaefer et al. 2017). Again, in Figure 5.7a for HC we can see high connectivity score between regions of the same network. We also see connectivity between regions of same network across left and right side of the brain. The diagonals on top and bottom of the main diagonal shows this. Whereas the DC of SZ subjects is weakly connected compared to HC and is mostly shows intra-network connectivity. The sparsity explains and support the existing literature explaining SZ as functional dysconnectivity between brain networks (Culbreth et al. 2021; Yu et al. 2011; Zhang et al. 2019; Zhu et al. 2020; Morgan et al. 2020b; Lynall et al. 2010; van den Heuvel et al. 2010).

Figure 5.7b compares male and female groups based on region-based (ROIs) HCP data. We can see similar patterns of hyper-connectivity of DMN and hypo-connectivity of SMN in females as compared to males. As the region-based (ROIs) parcellation divides the brain into left and right, we also see that females have high intra-network connectivity between left and right side of the brain as compared to males.



(a) FBIRN DC

(b) HCP DC

Figure 5.7 The figure compares the estimated DCs of HC with SZ and male with female using region-based (ROIs) FBIRN and HCP data. 5.7a show high weakly connected brain networks for SZ subjects whereas 5.7b show hyper-connectivity of DMN and hypo-connectivity for SMN for females as compared to females. The black and grey color denotes the regions in left and right side of the brain. Refer to Table 5.6 for a statistical comparison between female and male DCs.

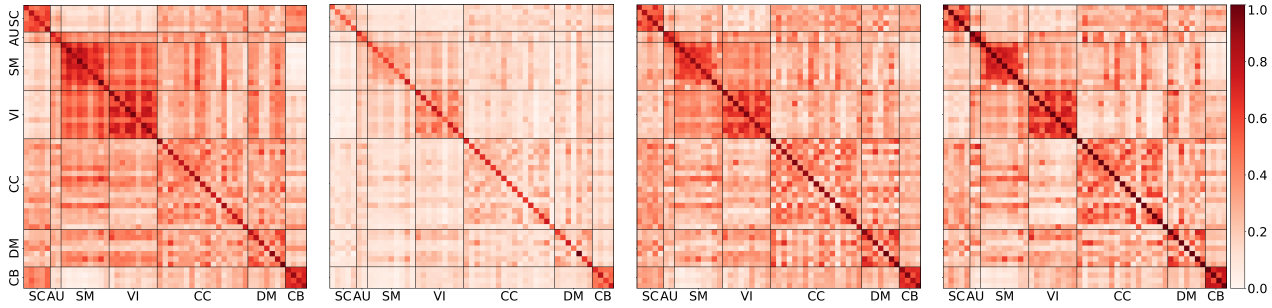
To verify the visual results, statistical testing is used to compare the DMN and SMN between males and females. The stats confirm the visual results with 1) female DMN showing higher connectivity than female SMN and male DMN, and 2) male SMN showing higher connectivity than male DMN and female SMN. We also see that the networks are highly statistically different. Refer to Table 5.6.

Table 5.6: Shows stats between male and female DCs (5.7b) estimated using region-based (ROIs) HCP dataset. We clearly see that females have hyper-connectivity in DMN and hypo-connectivity in SMN as compare to males. Female group has higher connectivity scores in DMN compared to SMN and male DMN whereas male group has higher connectivity in SMN compared to DMN and female SMN. This shows that our learned model accurately captures the differences in DMN and SMN connectivity among males and females and uses that for classification. F - Female, M - Male, L - Left, R - Right. Table 5.5 shows the p-value significance ranges.

Network 1	Network 2	Test Type	P-value	Avg. Connectivity 1	Avg. Connectivity 2
F_All	M_All	t-test	1e-14	0.455	0.533
		manwhitneyu	1e-25		
F_L_DM_temp	F_L_SM	t-test	2e-3	0.689	0.632
		manwhitneyu	4e-3		
F_R_DM_temp	F_R_SM	t-test	7e-4	0.671	0.593
		manwhitneyu	4e-4		
M_L_DM_temp	M_L_SM	t-test	2e-7	0.567	0.622
		manwhitneyu	1e-3		
M_R_DM_temp	M_R_SM	t-test	9e-4	0.558	0.611
		manwhitneyu	2e-4		
F_L_DM_temp	M_L_DM_temp	t-test	4e-5	0.689	0.567
		manwhitneyu	6e-5		
F_R_DM_temp	M_R_DM_temp	t-test	8e-5	0.671	0.558
		manwhitneyu	3e-5		
F_L_DM_pCunPCC	F_L_SM	t-test	2e-4	0.718	0.632
		manwhitneyu	1e-3		
F_R_DM_pCunPCC	F_R_SM	t-test	1e-5	0.758	0.593
		manwhitneyu	5e-5		
M_L_DM_pCunPCC	M_L_SM	t-test	2e-7	0.548	0.622
		manwhitneyu	3e-4		
M_R_DM_pCunPCC	M_R_SM	t-test	1e-2	0.547	0.611
		manwhitneyu	1e-2		
F_L_DM_pCunPCC	M_L_DM_pCunPCC	t-test	2e-4	0.718	0.548
		manwhitneyu	3e-4		
F_R_DM_pCunPCC	M_R_DM_pCunPCC	t-test	3e-4	0.758	0.547
		manwhitneyu	7e-4		
F_L_SM	M_L_SM	t-test	1e-1	0.632	0.622
		manwhitneyu	4e-1		
F_R_SM	M_R_SM	t-test	1e-2	0.593	0.611
		manwhitneyu	2e-3		

5.2.2.4 Task dependent DNC

Human brain can be divided into multiple parts/regions where each region is linked with a set of tasks. For example, the hippocampus is associated with memory. Thus it is important to know which region/network(s) are linked with the downstream task (e.g. disorder classification). Finding the linked regions/networks would help understand the disorder and allow to study the association of these regions/network(s) with the disorder in more detail. This section shows how the DNC structure learned by DICE changes and identifies different networks for the same subjects based on the downstream task. For this purpose, this study performs an experiment, where the estimated DNC for OASIS data is compared when predicting dementia, age and gender of the same subjects. The number of subjects were balanced with both HC and patients equalling 50% of the total subjects but had $\sim 62\%$ female subjects. Figure 5.8 shows that DICE produces task dependent DNC and the networks/domains showing high connectivity for each task adheres to the existing literature. The Figure 5.8a shows the DNC learned when classifying subjects for dementia. We can see high connectivity for components in the SM, DM, and CB networks. These networks are linked with dementia in existing literature, which support the results of the proposed method. Whereas when classifying gender of same subjects, the estimated DNC is different and show high connectivity for components in DM and reduced connectivity for SMN. Figure 5.8d shows the FNC computed via PCC for the same subjects. As FNC computed using PCC is only data dependent, the FNC would remain same for all the tasks and shows the inflexibility of the method. Figure 5.8 therefore shows a) DICE learns task dependent DNC and b) DICE accurately finds networks linked with the downstream classification task. This as a significant advantage over studies which compute a fixed/static FNC using PCC and hence is independent of the downstream task. We can see that Figure 5.8b which is the learned connectivity structure when predicting age does not show high connectivity between networks and the connectivity values for SMN and DMN are almost same. This could be a reason of small age variance in the dataset. Statistical scores are used to verify the visual results. Table 5.7 shows the statistical difference between the three DCs as a whole and between DMN and SMN. The estimated DCs are also compared with FC 5.8d.



(a) Dementia DNC (b) Age DNC (c) Gender DNC (d) PCC FNC

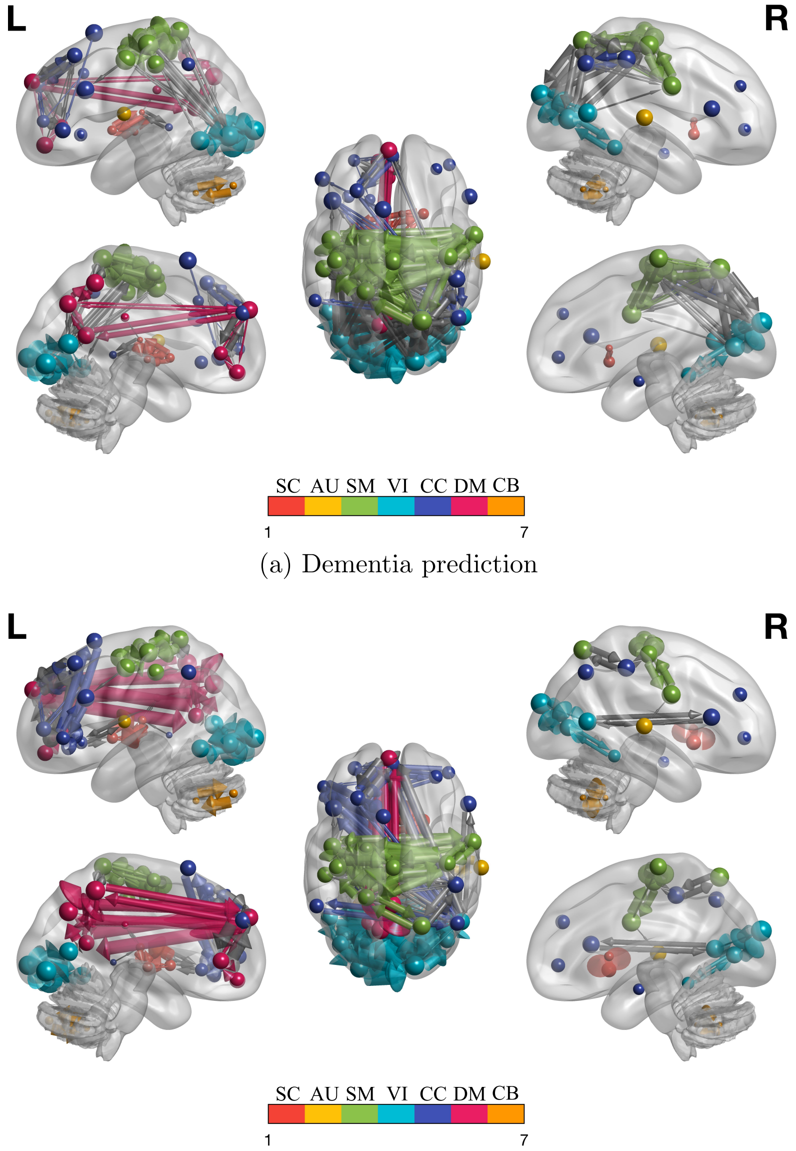
Figure 5.8 The figure shows how DICE estimates flexible DNC structures based on the ground-truth signal. DICE was trained for different classification tasks and use same test subjects to compare the estimated DNC for the subjects. All figures are mean DNC estimated for the same subjects with 5 randomly-seeded trials. 5.8a is the mean connectivity matrix estimated by DICE when trained to classify dementia. We can see high connectivity values for SC, SM, and CB networks. 5.8c is the mean DNC for the same subjects when the model is trained for gender prediction. We notice lower SM network connectivity and higher connectivity for DM network when predicting gender of OASIS subjects. 5.8d is the FNC computed using PCC. The FNC is independent of the task and would remain fixed (inflexible).

Table 5.7: Statistical difference of the learned connectivity matrices for OASIS ICA when predicting dementia, age and gender. The results show that the learned connectivity matrices are highly statistically different and SMN gets higher connectivity scores than DMN for dementia prediction whereas the opposite is seen for gender prediction.

Network 1	Network 2	Test Type	P-value	Avg. Connectivity 1	Avg. Connectivity 2
Dementia_All	Age_All	t-test	5e-22	0.323	0.168
		manwhitneyu	1e-38		
Dementia_All	Gender_All	t-test	2e-3	0.323	0.311
		manwhitneyu	8e-4		
Age	Gender	t-test	2e-301	0.168	0.311
		manwhitneyu	1e-301		
Dementia_DM	Dementia_SM	t-test	1e-7	0.478	0.645
		manwhitneyu	8e-8		
Age_DM	Age_SM	t-test	6e-1	0.294	0.308
		manwhitneyu	6e-2		
Gender_DM	Gender_SM	t-test	4e-1	0.527	0.555
		manwhitneyu	1e-1		
FNC_DM	FNC_SM	t-test	3e-2	0.487	0.580
		manwhitneyu	7e-3		
Dementia_DM	Age_DM	t-test	9e-6	0.478	0.294
		manwhitneyu	5e-7		
Dementia_DM	Gender_DM	t-test	2e-1	0.478	0.527
		manwhitneyu	1e-1		
Age_DM	Gender_DM	t-test	3e-7	0.294	0.527
		manwhitneyu	5e-8		
Dementia_SM	Age_SM	t-test	8e-34	0.645	0.308
		manwhitneyu	3e-23		
Dementia_SM	Gender_SM	t-test	4e-4	0.645	0.555
		manwhitneyu	1e-4		
Age_SM	Gender_SM	t-test	1e-18	0.308	0.555
		manwhitneyu	4e-17		

We can see that all three DNCs are extremely statistically different. It is also proven that DMN is given higher connectivity scores for gender prediction whereas, SMN connectivity is much higher when predicting dementia comparing to gender and age prediction tasks. To clear how the connectivity values change for DMN and SMN this study point outs the average connectivity scores of the networks for dementia and gender classification and compare it with the values of DMN and SMN computed via PCC. The connectivity values in FC for SMN and DMN are 0.580 and 0.487 respectively (and would remain same irrespective of the classification task). Whereas, when classifying dementia DICE show much higher SMN average value of 0.64 and a little decreased value of 0.478 for DMN showing a focus on SMN despite having more female subjects in the test set. When predicting gender for the same subjects the DNC estimated by DICE has a decreased SMN value of 0.555 and increased value of 0.527 for DMN hence focusing less on SMN and more on DMN when compared to the dementia classifying task thus verifying that the estimated DCs are task-dependent and not only data dependent. The meaning and significance of this result are further discussed in section 5.3.3.

To see the matrices as graph of nodes (regions) and edges (connectivity), Figure 5.8a and 5.8c are plotted on the brain and show the results in Figure 5.9. The figure shows high number of nodes and edges among components of VIN and SMN and among the two networks for dementia classification 5.9a, and high number of nodes and edges among components in DMN for gender classification 5.9b.



(a) Dementia prediction

(b) Gender prediction

Figure 5.9 The nodes and top 10% edges of the DCs are mapped on the brain, estimated for dementia and gender classification tasks, performed on OASIS dataset (same subjects). The size of the nodes is the sum of the outgoing and incoming edge weights. The arrows shows the direction of connectivity. We can see a high number and size of nodes and edges for SMN and VIN for dementia 5.9a, whereas for gender 5.9b we can see high node and edge size for DMN. Compare the red (DM) nodes and edges in Figure 5.9a with Figure 5.9b in the left side figures. Figure 5.9a also shows high connectivity between SM and VI networks which is missing in Figure 5.9b (right side figures). This reveals the networks and edges (graphs and subgraphs) relevant to the classification signal (e.g disorder) without need of comparison with other data. The results and their impact are further discussed in section 5.3.3.

Table 5.8: Dementia, gender classification and age prediction results on OASIS dataset. The table shows the DICE's results with ML methods using FC computed via PCC. Even with hand-crafted features ML methods perform similarly as our model. It is possible that the same input because of FC being only data dependent is one of the reasons of ML methods performing lower than DICE for Dementia and age prediction.

Dataset	Model	Task	N_Folds	Input	Metric	Score
OASIS	DICE	Dementia classification	10	ICA	AUC	**0.752**
OASIS	Logistic Regression	Dementia classification	10	FNC	AUC	0.745
OASIS	DICE	Gender classification	10	ICA	AUC	0.906
OASIS	Logistic Regression	Gender classification	10	FNC	AUC	**0.948**
OASIS	DICE	Age prediction	10	ICA	MAE	**6.14**
OASIS	Linear Regression	Age prediction	10	FNC	MAE	7.17
OASIS	Lasso	Age prediction	10	FNC	MAE	6.89

5.2.2.5 Dynamic Connectivity States

Studies like (Sakoğlu et al. 2010; Allen et al. 2012; Hutchison et al. 2013; Calhoun et al. 2014) show that human's brain FC is dynamic and can be used to find patterns which are not visible in static FC studies. These studies show that dynamic FC show re-occuring patterns. To study these patterns, dynamic connectivity of the human brain is divided into distinct k states (Rashid et al. 2014; Damaraju et al. 2014; Fu et al. 2021a). There are multiple methods proposed to find the k states with k-means being one of the most used methods. These studies show that the transition and time spent in each state is different for patients (SZ, dementia, autism) and HC. To validate the acquired results and to find such patterns k-means is used to find k (5) such states using the DCs estimated by DICE for FBIRN dataset. The time spent by both groups (SZ and HC) per state is calculated and compared.

Figure 5.10 shows that SZ subjects spend more time in weakly connected states (1,3) than HC which stay in states which show high connectivity score for visual (VI) and sensorimotor (SM). We also see that HC tend to change state more often than SZ which spend $\sim 66\%$ time in one state (number 3). Existing studies (Yaesoubi et al. 2018; Miller & Calhoun 2020b,a) show that window-less approach can find dynamic patterns that are not captured by the vastly used window-based approach. As DICE is an instantaneous model, this study investigates if DICE can capture more dynamic states than the window-based dynamic-FNC studies. For this purpose, using elbow method (Marutho et al. 2018), it is found that the best k for the estimated DCs is not 5, and set $k = 10$ and show the resultant states in Figure 5.11.

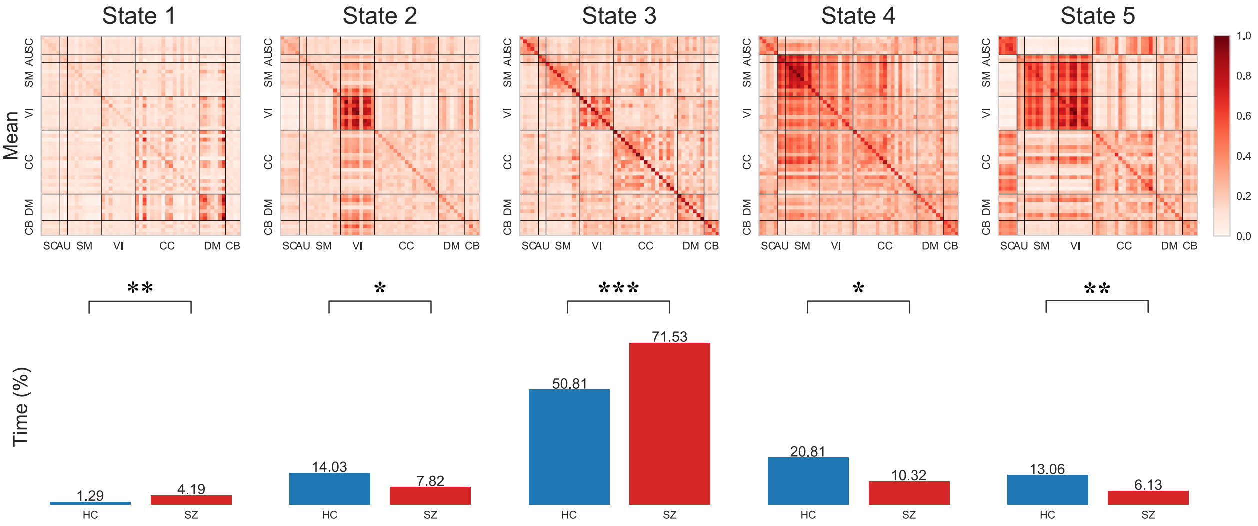


Figure 5.10 Five states computed using k-means on the DCs estimated by DICE for FBIRN dataset. First row shows the k means of the estimated DCs, second row shows the percentage time spent by both groups in each state, with the total time points being 155. Time spent in each state by SZ and HC differ significantly and matches the existing literature. The figure shows that a) time spent in each state is different by HC and SZ, b) SZ spend much more time in state 3 (weakly connected) than HC, c) HC spend more time than SZ in states (2,4, 5) which show high connectivity for VI, and SM networks, and d) Standard deviation of time for SZ is much higher (320.47) than HC (206.26) which shows that SZ stay in one state much more than HC which tend to change state more often. The stars denote the significance of difference in time spent in each state by the two groups. Table 5.5 shows the p-value significance ranges.

We can see that the model captures additional states that were not visible with $k = 5$. The additional states found show the pattern of directionality, specially in the states where HC spend more time than SZ. For example, in Figure 5.10, state 2 show dense connectivity for components in VIN and the direction is from VI to other states, and state 5 show similar direction but with sparse connectivity. Figure 5.11 captures the additional state (9) which shows the opposite direction, that is, VIN has mostly incoming edges. Presumably, this state represents the brain activity when different networks (e.g. SMN) are giving input to VIN to control the vision. This result is discussed in section 5.3.4.

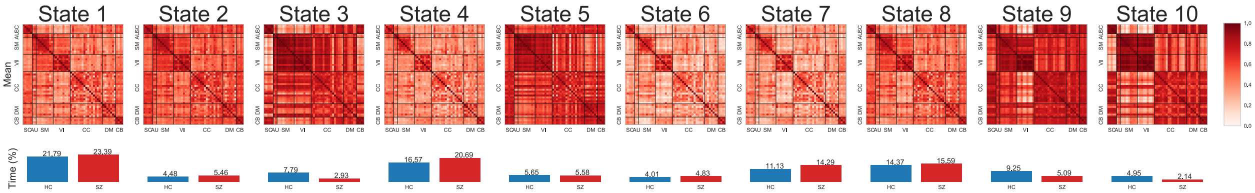


Figure 5.11 *10* states captured by k-means on the temporal DCs estimated by DICE on FBIRN complete dataset are shown here. The rows shows the means and the percentage of time spent by HC and SZ subjects in each state. DICE can capture more states than the standard (4-5) states captured by window-based approaches. The additional states not present in Figure 5.10 show the change of direction in connectivity. State 9 shows the opposite direction of connectivity between VIN and other networks, where VIN has mostly incoming edges. The ratio of time spent by HC and SZ subject in different states is similar to the results of Figure 5.10.

5.2.3 Temporal Attention

The proposed temporal attention module finds the important time-points that are relevant for the downstream task (e.g. gender prediction). As not all time-points are equally important for the downstream task, and fMRI data has low temporal resolution, the temporal attention is an effective way of finding important bio-markers for neuroimaging dataset. Finding the relevant time-points can help reduce the data and allow to focus on activities at specific points. Figure 5.12 shows the weights assigned to the subjects of FBIRN.

Weights for 16 subjects (8 per class) with 10 randomly-seeded trials are shown. The results show that the temporal attention module is very stable and assign similar weights to the time-points for every trial.

To further check the correctness of the time-points selected by DICE and how these time-points are useful in terms of classification performance, this study performs an experiment

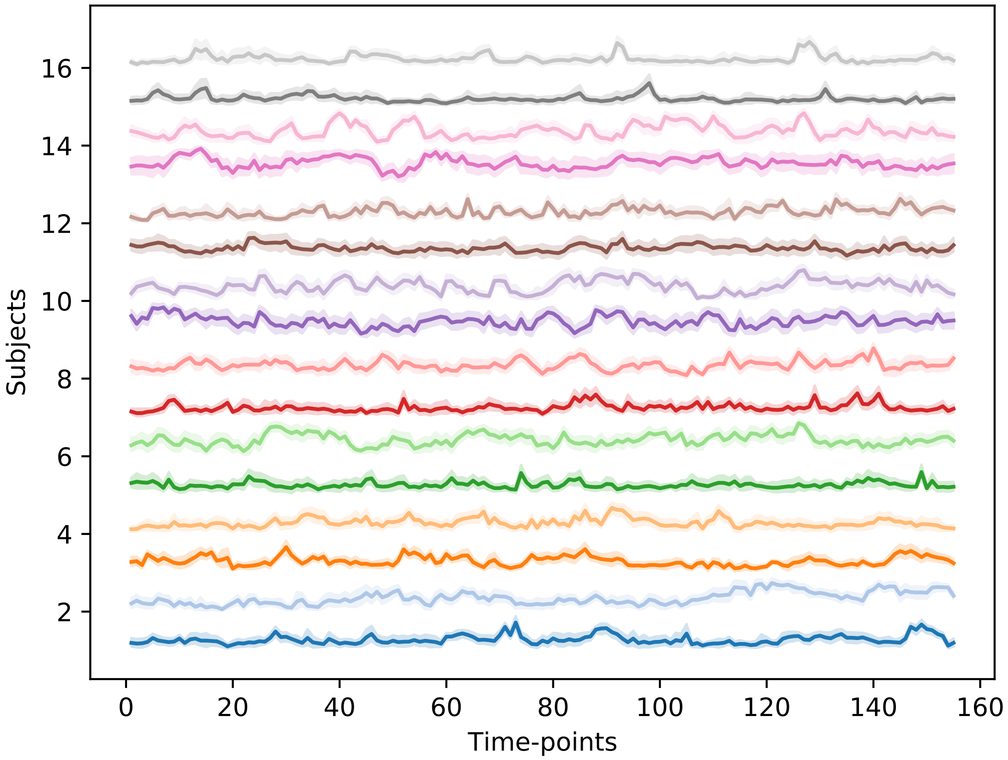


Figure 5.12 Temporal Attention weights for one of the test folds (16 subjects) of FBIRN. Attention weights are computed using GTA module. X and y axis represent time-points and subject number respectively. The figure shows that for each subject, the attention weights remain stable across multiple randomly-seeded trials (10). The values of the 10 trials are used to create the confidence interval for each subject. The consistency is greatly increased with an increase in number of training subjects. **Note:** For each subject the subject number was added to the attention weights to separate the weights, as for each subject the weights have a range of $0 - 1$. Dark and light colors represent SZ and HC subjects respectively.

where after training the model, $\mathbf{W}_t$ of the top 5% values was used to train an LR model and then use the top 5% time-points of the test data to test the model. Similar experiments for bottom 5% values were performed as well. Table 5.9 shows the comparison for the three brain disorder dataset. The results show that the LR model provides high AUC score by just using 5% of the important time-points. Thus, it proves that a) not all time-points are important for classification of the downstream task and b) DICE accurately finds the important time-points. This study uses an LR model for this experiment to show that the learned top and bottom 5% values are not limited to the proposed DICE model but is generalized such that an independent LR module gives high classification performance using the top 5% data identified by DICE and does not learn on the low 5% data. Finally, the experiments also show that not using the temporal attention reduces the model classification performance by upto 10% A.2.

Table 5.9: AUC score comparison on brain datasets with ICA components by using all, top 5% and bottom 5% time-points only. A logistic regression (LR) model was trained using the time-points identified by DICE and compare the results when using top and bottom 5% time-points. We can see that using only top 5% time-points are enough to almost reach the AUC using all time-points.

	Method	FBIRN	OASIS	ABIDE
100%	DICE	0.86	0.752	0.722
Top 5%	LR	0.85	0.743	0.706
Bottom 5%	LR	0.566	0.548	0.532

5.3 Discussion

The experiment of this study revealed a number of interesting properties of DICE and uncovered some interpretable directed connectivity graphs that are of high utility for the neuroimaging field. As supported by results, models with glass-box layer like DICE have a high potential for studying resting-state dynamics of the brain. In the following sections most pertinent results are discussed.

5.3.1 Inter-network and Directed Connectivity

Results in Sections 5.2.2.1 and 5.2.2.2 show that DICE infers DNC that agrees with the essential findings of the FC studies (Yan et al. 2017; Parisot et al. 2018; Kawahara et al. 2016; Ktena et al. 2017; Arslan et al. 2018; Kazi et al. 2021; Kim & Ye 2020; Ktena et al. 2018; Ma et al. 2019) and provides two additional aspects: inter-network connectivity and direction of connectivity. The inter-network connectivity is of great significance as the brain is not made up of isolated networks and many tasks require information passing and neurons firing through multiple networks. Thus making it crucial to find how these networks are connected to each other if connected at all for patients and controls. Capturing the dysconnectivity between networks for patients can lead to knowledge discovery about the functionality of the human brain and the effects of brain disorders on it. Furthermore, finding directionality between networks is also of great significance. This study showed in experiments that DICE captures the direction of connectivity between networks. The direction of connectivity from VI to other networks, and from CC to SM networks is justifiable. Existing studies (Breukelaar et al. 2017; Cole & Schneider 2007; Tsai et al. 2019) show that cognitive control is responsible for functions like attention, remembering, and execution. These functions are often required when doing a motor task controlled by sensorimotor, which hints at the direct effect of the CC network on the SM network, captured by DICE. Regarding VI and other networks, it is known that VI is mostly a means of input (visuals) to our brain, which is then processed by different parts of the brain. Thus, most of the flow of information is from VI to other networks and few in the opposite direction, which is required to control VI for accomplishing different motor tasks controlled by SM. Therefore, the experiments also show that most incoming connections to VI are through the SM network, thus accurately capturing the flow of information between networks. This flow of information is not captured in simple correlations. These two aspects can be crucial in understanding brain working and are currently missed in connectivity estimation methods such as FNC.

Directed connectivity directed influence of an intrinsic brain network on other networks. Estimating the direction of connectivity may simplify targeted interventions that are instrumental in establishing causal relations. Capturing causality between networks further helps to understand complex systems and answer counter-factual questions (Schölkopf et al. 2021), and is left to future work. The proposed model finds non-negative relations between components/nodes, which are considered as dependencies or relevance rather than correlations. However, the negative correlations in FC and FNC are also helpful and provide descriptive information. It might be an easy fix to incorporate negative relations in connectivity matrices estimated by DICE. This is discussed in section B.

5.3.2 Interpretability

Section 5.2.2.3 shows how the DC and DNC estimated by DICE are interpretable in how accurately they capture the difference in connectivity between 1) schizophrenia patients and controls and, 2) male and female groups. In classifying schizophrenia patients from controls, DICE learned the most significant differences were in the VI, SM, and DM networks. Controls show robust connectivity of VI and SM with each other and with other networks, which is missing for SZ patients. The finding of dysconnectivity and/or lower connectivity scores for VI and SM networks for SZ patients is not surprising as there exists ample evidence in prior studies of schizophrenia leading to multiple abnormalities related to visual and motor functions such as perception of contrast and motion, detection of visual contours, and control of eye movements to name a few (Silverstein & Rosen 2015; Butler et al. 2008; Chen et al. 1999; Kéri et al. 2002). These abnormalities certainly affect motor skills which presumably is a reason for the low connectivity for SM and VI networks captured by DICE for SZ patients. DICE also captures hyper-connectivity in DMN for SZ patients which is reported by existing studies (Guo et al. 2017).

Whereas in classifying gender in the same dataset, DICE emphasized hyper-connectivity in the DM network and hypo-connectivity for the SM network for females compared to males. The differences captured in the DC and DNC for both tasks are supported by existing studies (Culbreth et al. 2021; Yu et al. 2011; Zhang et al. 2019; Zhu et al. 2020; Morgan et al. 2020b; Lynall et al. 2010; van den Heuvel et al. 2010; Kim et al. 2021; Filippi et al. 2013; Mak et al. 2016; Ritchie et al. 2018) that show the role of the DMN in gender classification and VI dysconnectivity for schizophrenic patients. Similarly to existing studies (Zhang et al. 2018b; Ingalhalikar et al. 2014), DICE shows that female subjects have higher connectivity between the contralateral homologue brain networks relative to males.

DL models are commonly viewed as black-box models because of the difficulty of interpretation and not easily explained performance on the tasks they are trained on. These models can show excellent performance on tasks such as classification based on the reasons that are not substantially revealing about the input data nor their dynamics. One reason is shortcut learning (Geirhos et al. 2020): a DL model can classify images with or without airplanes with high accuracy by paying attention exclusively to the background (blue sky). Although predictive, such models cannot help in knowledge discovery. To control for shortcut learning it is important to be able to see why predictions are made. One approach is making DL model interpretable. For that a posthoc method is often used, e.g., saliency maps (Simonyan et al. 2014; Ras et al. 2021; Angelov et al. 2021; Lewis et al. 2021). Such methods explain the input data by finding which part(s) of the input the model is most sensitive to. Saliency maps have shown some good results in computer vision tasks in 2d images. The use of saliency maps in neuroimaging and temporal data has different challenges (Liu et al. 2021) as the output maps are noisy, difficult to interpret and does not provide good boundaries nor the connection between different salient regions. Selection of the method for obtaining saliency maps is also something to consider as some of the methods are architecture based. Hence, using saliency maps to get task-specific brain's connectivity graph is not feasible using current methods. To overcome the black-box nature of DL models and avoid using a posthoc method, this study focused on the interpretability of the model's results. For this purpose, as brain disorders are commonly associated with disruptions in the connectivity pattern of brain networks, this study used only the learned connectivity matrices by DICE for the downstream classification or prediction tasks, thus making the model extract the abnormality in connectivity relevant to the ground-truth signal. One way to conceptualize about the proposed approach is to think of the generated DC and DNC as a “glass-box layer” (clear and interpretable) layer as noted in Figure 5.1. This approach combines flexibility (the layer is trainable) with interpretability and enables the model to capture differences in the connectivity of the groups in classification task. Regression is also possible with the proposed approach, although it is left for the future work. The proposed “glass-box layer” approach enables learning the essential networks and their connection to other networks relevant to the training signal and directly output that without using a posthoc method. As the DC and DNCs estimated by DICE are based on learnable functions, the output matrices can have slightly different values when the model is retrained, which is an attribute of DL models. Therefore, all the connectivity matrices shown in the chapter are averaged over several randomly-seeded trials.

5.3.3 Task-dependent Flexible DNC

This work fully utilizes the flexibility of the proposed DL model to learn task-dependent (ground-truth signal) directed connectivity structures. It is shown in Section 5.2.2.4 that DICE estimates DNC structures for the same subjects that are distinct to the ground-truth task of dementia, age, or gender. Hence DICE can show the networks and their connectivity crucial for specific downstream tasks. The networks identified by the model through the learned DNC for dementia classification (SM, CB, VI) match the results of prior studies (Ingalhalikar et al. 2014; Albers et al. 2015a; Filippi et al. 2017; Grant et al. 2014; Jacobs et al. 2017). Whereas, for gender prediction, the most prominent network identified by the network was DM, which again matches existing literature (Kim et al. 2021; Filippi et al. 2013;

Mak et al. 2016; Ritchie et al. 2018). This is a strong validation of the ability of DICE to find disorder-dependent networks and connectivity patterns. This study showed in Figure 5.8a that DICE focused more on SMN than DMN despite having almost two-thirds of female subjects in the test set. This result is significant because the model learned that the SMN connectivity, is more important than DMN for the downstream task of dementia classification and hence enhances the signals for SMN. This eliminates the need to acquire strictly matched subjects with only the difference(s) for which you want to find the relevant networks and connectivity. For example, when trying to find the networks related to schizophrenia using PCC, one needs to find two groups (schizophrenia patients and controls) that do not have any other differences. Extraneous differences would create ambiguity regarding whether the networks identified are related to the disorder (schizophrenia) or some other difference, e.g., gender. Instead of explicitly confronting the confounding factors by regressing them out or taking equivalent measures, DICE performs the “de-confounding” implicitly based on the training labels.

Another notable property of DICE is that it finds the relevant networks and the connectivity structures (sub-graphs) without receiving them during training, making DICE a self-supervised graph learning model.

5.3.4 Dynamic DNC and Temporal-attention

As hypothesized, and shown in previous studies (Sakoğlu et al. 2010; Allen et al. 2012; Hutchison et al. 2013; Calhoun et al. 2014) results in Section 5.2.2.5 show that connectivity between brain’s intrinsic network is dynamic, and dynamic connectivity can capture patterns which are missed by static models. Notably, controls and SZ patients spend different amounts of time in each state 5.10. Controls spend more time than SZ patients in strongly connected states, especially for visual and sensorimotor networks. On the other hand, SZ patients spend time in weakly connected states and do not often spend time in other states. Similar patterns were observed in FNC studies (Rashid et al. 2014; Damaraju et al. 2014; Rabany et al. 2019; Wang et al. 2014; Yang et al. 2022).

Moreover, using all subjects in the FBIRN (Keator et al. 2016), DICE finds additional states doubling the state resolution. This temporal resolution increase is explained by instantaneity of directed connectivity estimation in DICE in contrast to using a sliding window. Therefore, estimating connectivity instantaneously makes the model robust and finds patterns that are missed when using a window-based approach. Another explanation and an additional factor is the increased richness of representation via a directed graph - the connectivity matrices of DICE have twice the number of parameters compared to FC and FNC. The experiment with *k=10* states show similar patterns of strongly and weakly connected states but they now vary in the direction of the connectivity. This result shows that both the connectivity strength and direction of connectivity are dynamic (changes over time). As this state is rare (based on time spent), it would be harder for window-based approaches to capture it. It would be interesting to see when and how the direction of connectivity changes and how external factors like performing a task can trigger these changes. This, however, is a topic of the future work.

Finally, this study showed that not all time-points of the fMRI data are equally important to the downstream prediction task and discriminative connectivity matrices exhibit temporal dynamics. Using temporal attention, DICE finds important time-points relevant to the ground-truth signal used in training. This further helps in interpretability as DICE finds the time-points where the brain activity shows signals relevant to the task. Potentially, this would also be important in task data where the subject is asked to perform different tasks, and the DICE model can be used to find out which task revealed the symptoms of the underlying disorder. The experiments show that temporal attention assigns stable and consistent weights to time-points across different randomly-seeded tasks. It was also notice that a) just 5% of time-points are sufficient for achieving high classification performance and b) exclusion of temporal attention (assigning the same weight to every time-point) negatively affects classification performance. Consistent temporal attention values across randomly-seeded trials further strengthens the evidence of temporally dynamic discriminative DCs and the value of attention mechanism. As the experiments show, the proposed attention module is indeed reliable per the definitions and potential issues discussed by Jain et al. (Jain & Wallace 2019) and Wiereffe et al. (Wiegreffe & Pinter 2019). As a learnable method, DICE and other "glass-box layer" models need to be able to consistently across training runs assign temporal attention values and estimate connectivity between nodes, whereas inflexible methods computing correlations such as PCC do not have this property. In a way, flexibility of the learnable model comes with an additional requirement of stability of learned interpretations. Even though DICE model works well by showing high classification performance and assigning consistent self and temporal attention values on relatively small datasets, having more subjects for training leads to an even more consistent assignment of temporal weights.

5.4 Conclusions

The work in this chapter demonstrates importance of learnable interpretable estimators of dynamic, directed, and task-dependent connectivity graphs from fMRI data. DICE learns to estimate interpretable dynamic and directed graphs that represent the directed connectivity among brain networks. The end-to-end training process removes the need for existing external methods such as PCC and K-means, which are interpretable but inflexible and strictly depend on the input data. Implementing DICE with glass-box layer allowed to bypass the need for a posthoc method for interpreting learned model representations.

Connectivity matrices estimated by DICE show how brain connectivity changes across disorders, genders, and age. The learned connectivity matrices help understand the human brain and its disorders as the actual ground-truth connectivity matrix is not available. Furthermore, this study moved from FC and FNC to DC and DNC to learn the direction of connectivity and simultaneously removed the issue of window sizing of input data by making the model instantaneous. The learned connectivity matrices provide knowledge that adheres to existing studies. Utilizing flexibility of DL models in learning data representations, this study shows that using the same data, distinct connectivity structures can be learned based on the downstream task and the ground-truth signal. This flexibility allows acquiring more information from the data by using different training labels, which would require a much more involved process of data selection and manual filtering out of confounding factors for methods that are fully determined by the data, like PCC. DICE highlights different networks linked with the downstream classification task, e.g., the default mode network for gender prediction. Unlike other interpretable models that may pay for it with a decrease in classification performance (Johansson et al. 2011; Dhurandhar et al. 2018; Luo et al. 2019; Shukla & Tripathi 2012), DICE beats state of the art methods in multiple classification problems on four neuroimaging datasets.

For classification DICE uses the learned connectivity structures. Together with the temporal weights these structures are reasonably consistent across varying seeds. Notably, DICE's performance drops without the use of temporal attention. The temporal attention module of the model finds interpretable bio-markers crucial to performing the classification task and shows that only a small fraction of time-points is enough for attaining maximum performance. Notably, not all time points are discriminative, as evident from the sparse distribution of temporal attention weights in Figure 5.12 and high predictive power of just the top 5% of the attention weights of Table 5.9.

As the ground truth for the dynamic graph structure in resting state fMRI is unavailable, there is a need for models with "glass-box layer" like DICE that can estimate this structure based only on the data and classification labels.

Next, the goal is to use a self-attention based module on the temporal axis as well. This would allow to replace the 'black-box' biLSTM with a 'glass-box' module.

5.5 Spatio-temporal Self-attention

This section presents a glass-box transformer model *Glacier* that provides interpretability on spatial and temporal dimensions. Glacier shows that unlike the hybrid models deep learning can be successfully applied to neuroimaging without incorporating another method. Glacier uses self-attention (Vaswani et al. 2017) and mixes space and time dimensions to create directed and dynamic connectivity matrices between brain's intrinsic networks.

The results show that the DNC matrices estimated by Glacier are downstream task-dependent and uncovers crucial spatial and temporal biomarkers. Results also show that using the estimated DNCs Glacier beats state-of-the-art models on brain disorder classification (schizophrenia, dementia, and autism), gender classification and age prediction.

5.5.1 Method

Traditionally, deep neural networks are used to create embeddings from the data. These embeddings created from different modules are used for downstream task but are often difficult to interpret. To solve this problem, this study presents *Glacier* as a deep learning model to estimate the dynamic connectivity graph between the nodes present in the dataset. Instead of the embeddings, Glacier uses only the learned graph structure for the downstream task. By not using the embeddings, the model is forced to learn distinct graph structure for the groups present in the data (e.g., HC and Patients). The idea of the model is to use self-attention to uncover task-dependent spatio-temporal dependencies while performing downstream classification. These dependencies are used to interpret the results and highlight important biomarkers. It can be noticed that interpreting distinct graphs are much easier than interpretting embeddings that are often in high dimension space. The directed edges of the learned graph represent the directed connectivity score estimated via self-attention. Glacier is composed of three main parts based on self-attention and attention. A temporal attention is performed to create a final static graph based on which classification is performed. The working of Glacier is explained in this section. Refer to Figure 5.13 for the architectural details of the model.

Temporal Mixing

Important information is present in the temporal dimension of datasets which are of sequential form (time-series). For example in speech datasets, the location of the words in a sentence have great significance and changing the order can result in different meaning. Similarly, temporal dimension of medical imaging data is also of critical nature, where indicator(s) of the interested problem are seen at specific time-points and which have a significance effect on other time-points. To capture these time-sensitive dependencies Glacier uses transformer encoder (Vaswani et al. 2017). Using the dependencies the encoder mixes the temporal dimension and outputs new embeddings which have more influence by the time-points that are identified as important by the model. These temporal dependencies are later used to interpret the temporal dimension.

Glacier is used to capture directed dependencies between every time-point. As the dependencies can be different for the nodes in the dataset, time-series of each node is given individually as input but share the weights of the model. To convert the scalar value x_t^i

representing the value of i^{th} node at time-point t into a vector, firstly a feed-forward neural network is used whose output is passed to the transformer encoder which outputs the vector c_t^i.

Spatial Mixing

Nodes of a system are dependant and affect each other through time. The rate of change in these dependencies is reliant on the underlying system. Activity in human brain is extremely dynamic and can change at any moment. To capture dynamic directed connectivity between nodes of the system, self-attention between the nodes at each time-point is used. At any time-point t the self-attention module receives N embeddings represented by vector $\boldsymbol{c}$ and output a depenedency matrix $\mathbf{W}$. The $\mathbf{W}$ matrix is represented as the directed connectivity between the nodes.

Temporal Attention and Classification

To create a single weight matrix $\mathbf{W}^f$ for downstream task the weighted average of $\mathbf{W}_{1-T}$ matrices are acquired. The same GTA module presented in Section 5.1.2.3 is used for that purpose.

5.5.2 Datasets and Training

Glacier is applied to neuroimaging datasets because of two major reasons. Firstly, neuroimaging is a field where results need to be transparent and interpretable and mere classification performance is not enough to trust the results. Secondly, as brain disorders are linked with

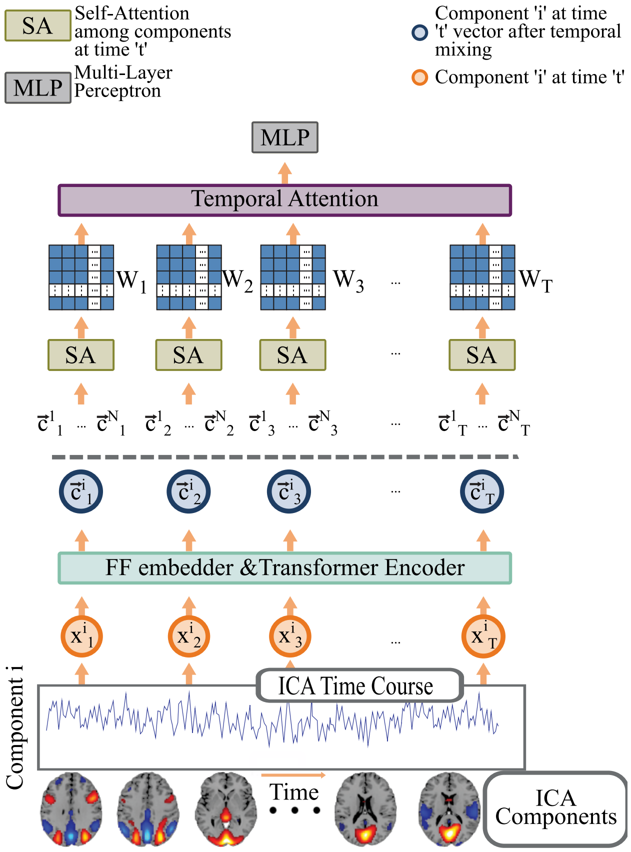

SA
Self-Attention among components at time 't'
MLP
Multi-Layer Perceptron
Component 'i' at time 't' vector after temporal mixing
Component 'i' at time 't'
MLP
Temporal Attention
W1
W2
W3
WT
SA
SA
SA
SA
FF embedder &Transformer Encoder
ICA Time Course
Component i
Time
ICA Components

Figure 5.13 Glacier is comprised of two self-attention modules for temporal and spatial mixing. Temporal attention on top is used for selecting important time-points. Multi-layer perceptron (MLP) take a final graph to make downstream classification.

dysconnectivity in the connectivity between brain's intrinsic network, Glacier which uses only the estimated graph between nodes for classification is an ideal fit for such fields and dataset. Glacier was tested on four different neuroimaging datasets. These datasets represent the functional activity of the brain captured via resting state functional magnetic resonance imaging (fMRI) scans. Four datasets used in this study include FBIRN (Function Biomedical Informatics Research Network[1]) (Keator et al. 2016) project, release 1.0 of ABIDE (Autism

[1] fBIRN phase III is used.

Brain Imaging Data Exchange[2]) (Di Martino et al. 2014) and release 3.0 of OASIS (Open Access Series of Imaging Studies[3]) (Rubin et al. 1998) to predict schizophrenia, autism and dementia respectively. Subjects from the ABCD(Adolescent Brain Cognitive Development [4]) Casey et al. (2018) datasets are used for gender prediction. FBIRN dataset was divided into 18 test folds were used all other datasets were divided into 10 test folds.

5.5.2.1 Preprocessing

Instead of using the voxel-level data, for disorder classification and gender prediction on ABCD data, independent component analysis (ICA) and for gender prediction on HCP predefined atlas based region of interest (ROIs) was used as brain parcellation method. Refer to Section 4.1.1 for details.

5.5.2.2 Training

Training of the model is performed similarly to as explained in Section 5.1.3 without some minor hyper-parameters changing. Mean area under curve - receiver operating characteristic (AUC-ROC) and other metrics are reported to show classification performance.

5.5.3 Results

This section shows the classification performance and the interpretable connectivity matrices estimated by Glacier.

[2] http://fcon_1000.projects.nitrc.org/indi/abide/

[3] https://www.oasis-brains.org/

[4] First scans from first session are used .

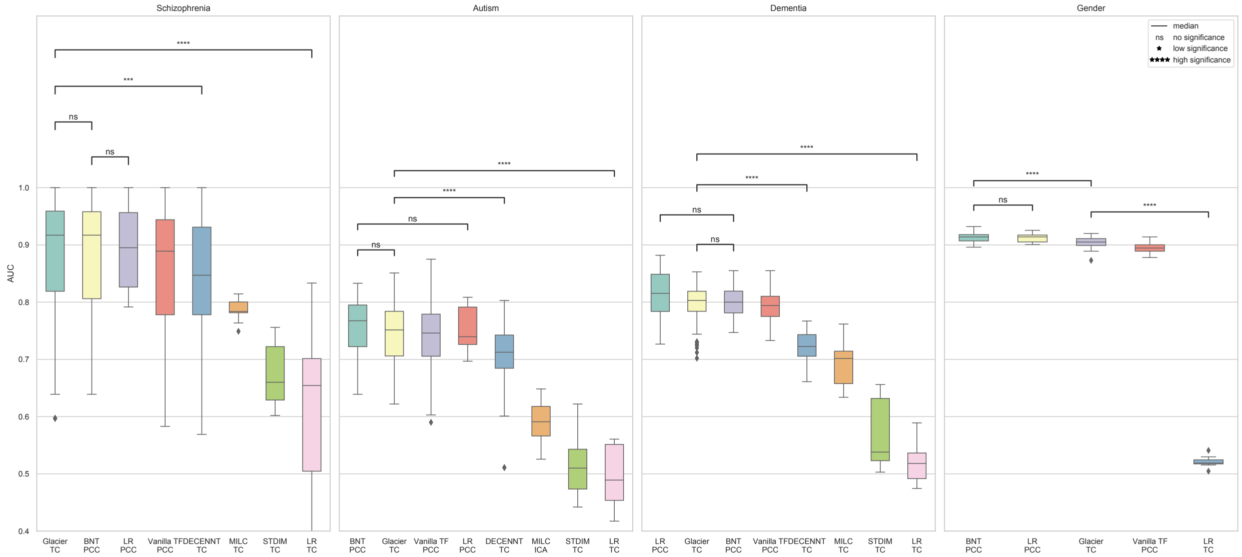


Figure 5.14 AUC comparision of Glassier model with six different methods (BNT (Kan et al. 2022), DECENNT (Mahmood et al. 2022) MILC (Mahmood et al. 2020a), STDIM (Mahmood et al. 2019a), LR, SVM), over three different datasets on ICA time courses. Glacier outperforms SOTA methods. It can be seen that when using ICA time-courses (TC) ML methods fail significantly, however SOTA DL and ML methods perform comparable to Glacier only if they are provided with hand-crafted features (FNC matrices computed by PCC.)

5.5.3.1 Classification

Glacier performs better or in reaching distance compared to state-of-the-art methods. 5.14 shows the AUC of Glacier for brain disorder classification using ICA time-courses as input data. 5.10 shows the performance for gender classification on HCP data using regions of interest (ROIs) extracted via Shaefer atlas (Schaefer et al. 2017).

Table 5.10: Classification performance comparison of *Glacier* with other DL methods on ROIs data of HCP. Our model gives comparable performance to state-of-the-art results in every metric. The best two scores are shown as bold and italic respectively. Note: The results for GCN Arslan et al. (2018) on HCP data are reported in GIN paper Kim & Ye (2020).

	Glacier	DICE	GIN	GCN
AUC	*0.935*	**0.936**	NA	NA
ACC(%)	*85.6*	**86.0**	84.6	83.98
Precision (%)	*85.3*	**87.2**	86.19	84.59
Recall (%)	**90.5**	*88.6*	86.81	*87.78*
Parcellation	Shaefer 200	Shaefer 200	Shaefer 400	Shaefer 400
Validation	10	10	10	10
Subjects	942	942	942	942

5.5.3.2 Interpretation

Group Differences

5.15 shows the ENC estimated by Glacier for HC and SZ patients using FBIRN dataset. The ENCs are the average of multiple test subjects using 10 randomly seeded trials. It is noticeable that HC show hyper-connectivity as compared to SZ patients. The inter-network dysconnectivity for SZ patients is reported in existing studies. Furthermore, HC show high connectivity between VI and SM networks and connectivity of subcortical (SC) network is shown with cerebellum (CB), default-mode (DM), and cognitive control (CC) networks.

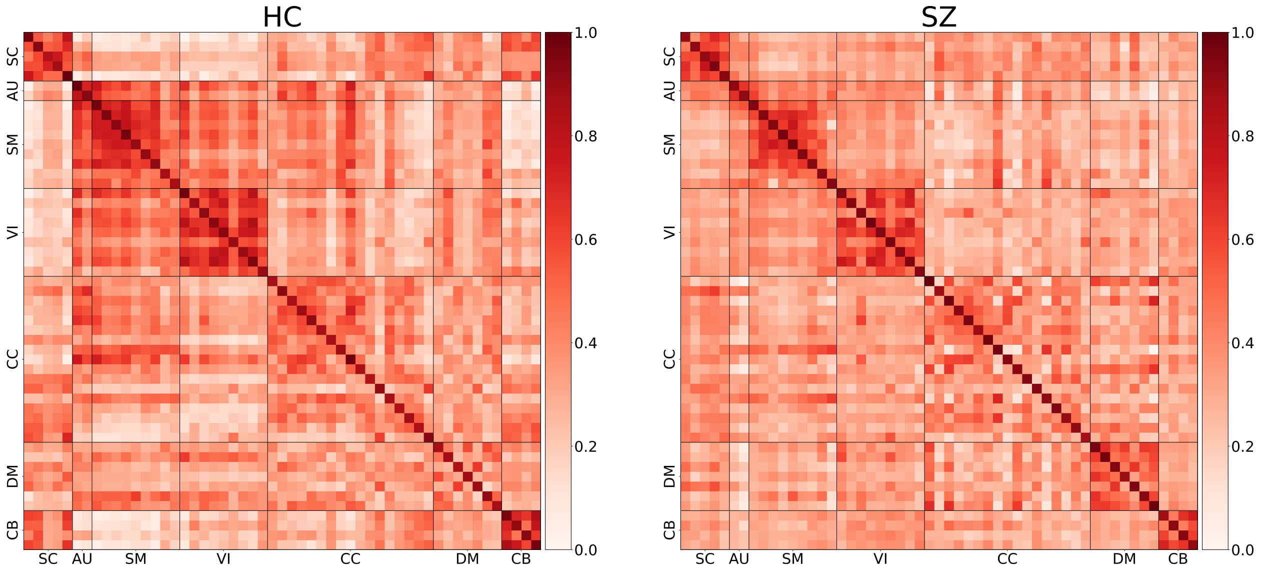


Figure 5.15 The estimated connectivity matrix of HC and SZ patients are compared. The axis show the components divided into 7 networks. Different connectivity patterns are shown for HC and SZ patients across multiple networks, especially VI, SM and DM networks.

Whereas, SZ patients does not show such patterns rather show hyper-connectivity in DMN as compared to the HC.

Flexible Connectivity Estimation

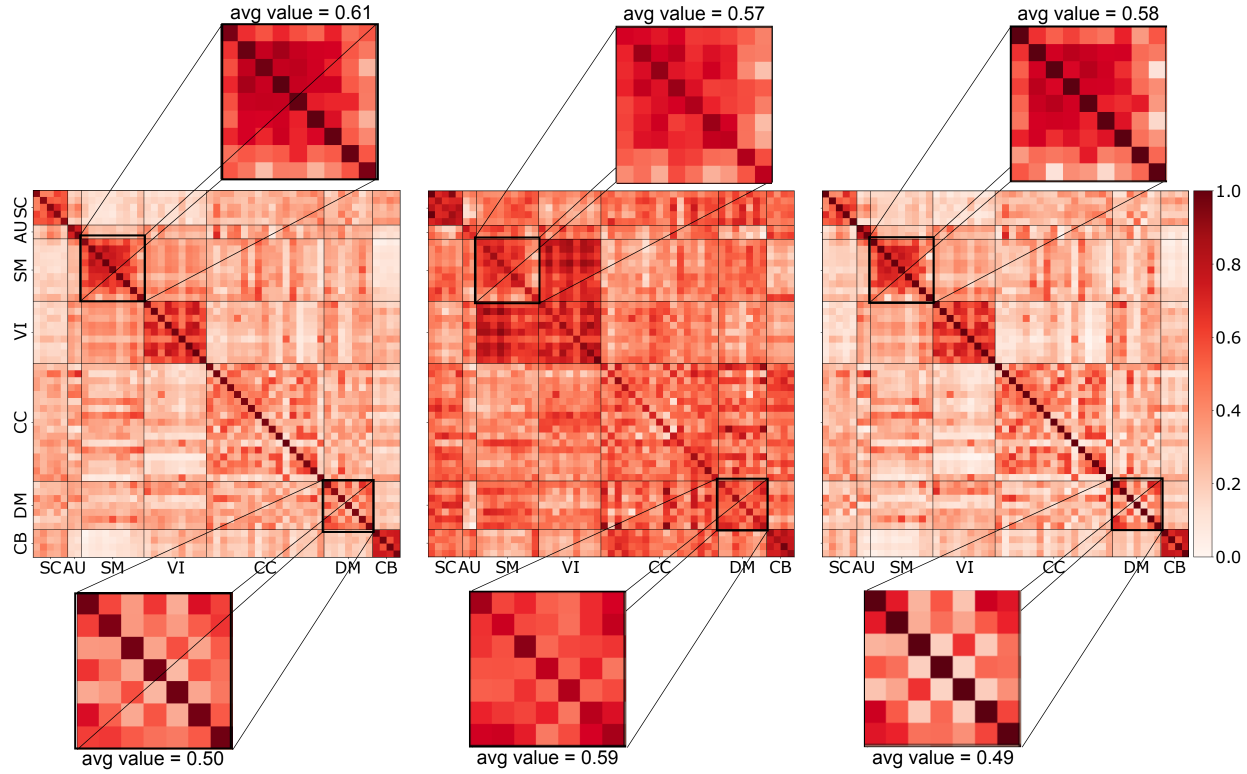
One of the biggest advantage of Glacier is that as the model extracts useful features (connectivity matrices) and does not rely on inflexible hand-crafted features. The flexibility allows Glacier to produce task-specific (e.g., brain disorder) connectivity matrix and produce subgraphs for a brain disorder. Thus highlights crucial brain networks connectivity patterns. Figure 5.16 shows how Glacier focuses on SM network for dementia prediction and on DM network for gender prediction of the same subjects. Static PCC based FC is inflexible and produces same result irrespective of the task.

Temporal Connectivity Estimation

As the proposed Glacier model also uses self-attention based module on the temporal axis, it is possible to visualise and interpret the connectivity weights estimated for the time-courses. Figure 5.17 shows the temporal connectivity matrices for the 7 networks for HC and SZ patients estimated on FBIRN dataset. We can see that for each network some of the time-points are assigned higher weights than the others.

5.5.4 Conclusions

Glacier shows the importance of using glass-box deep learning models that are interpretable. Glacier was used to estimate connectivity between brain's intrinsic networks. The estimated matrices not only provided high classification score but more importantly captured the differences in brain networks connectivity between HC and patients. This work also showed that deep-learning models can be successfully applied to fields like neuroimaging for clas-



(a) Dementia DNC

(b) Gender DNC

(c) PCC FNC

Figure 5.16 Glacier estimates flexible DNC structures based on the ground-truth signal. Glacier was trained for different classification tasks and use same test subjects to compare the estimated DNC for the subjects. 5.16a is the connectivity matrix estimated by Glacier when trained to classify dementia. 5.16b is the DNC for the same subjects when the model is trained for gender prediction. 5.16c is the FNC computed using PCC. The FNC is independent of the task and would remain fixed (inflexible). Notice how average value of SM in 5.16a is higher than 5.16b and 5.16c. Whereas average value of 5.16b is higher than 5.16a and 5.16c. This shows that Glacier gives more attention to disorder-specific networks. Importance of SM for dementia prediction and DM for gender predcition is shown is existing studies Albers et al. (2015b); Kim et al. (2021).

sification and interpretability without incorporating other statistical or ML methods. For future work, It would be interesting to test the model on task-fMRI data which would give the ground-truth important time-points depending on the task. This would allow to compare the time-points marked as important by the model against the ground-truth time-points.

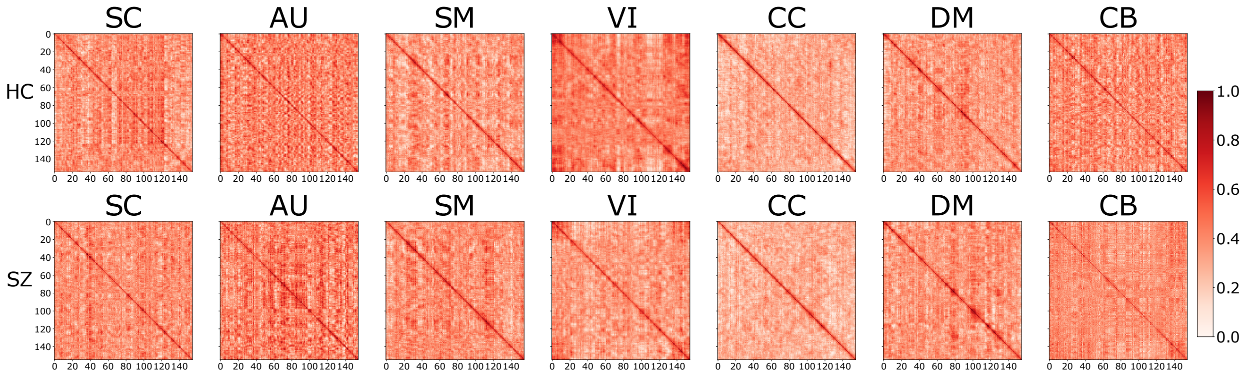


Figure 5.17 Temporal connectivity matrix estimated by Glacier for the time-points. The matrices of the components are divided into the same 7 networks and the average was taken of all the matrices in each network. The figure shows the the 7 matrices, one for each network for HC and SZ patients take from FBIRN dataset.

CHAPTER 6
CONCLUSIONS AND FUTURE WORK

This research in this dissertation showed that DL models could be instrumental in neuroimaging. Interpretable DL models can beat classical ML-based models without using hand-crafted features. More importantly, the flexibility of DL in learning representations allows for estimating subject and disorder-specific brain network connectivity graphs. The estimated graphs and subgraphs can be used to uncover disorder-specific biomarkers. The proposed idea of including the glass-box interpretable layer(s) significantly increases the interpretability of the representations learned by the DL model. This research is an initial step towards using DL to estimate the brain's network connectivity. Further research will take us closer to estimating the causality between brain networks.

For future work, a natural extension would be to omit pre-processing with a dimensionality reduction method—like ICA or region-based parcellation—and train a model end-to-end on the voxel-level data. This, however, may require substantially larger datasets and may not be as valuable as the current model for an average-sized research dataset. As the proposed models can estimate the direction of connectivity, for future work, it would be interesting to examine how the direction of connectivity changes through time and during tasks for HC and patients. A natural extension of this work would be the application on task-fMRI data and compare the temporal bio-markers with known important time-points.

Furthermore, implicit-layer(s) based methods, especially neural-ODE (NODE) should be a good next step of this study. NODE-based DL architectures have proven their value in fields where learning data dynamics is essential. NODE architectures applied on the temporal axis can help solve the problem of different temporal resolutions in datasets acquired from multiple sites, e.g., ABIDE. Learning the derivatives of the embeddings would also make it more interpretable and flexible, as shown in studies like (Hasani et al. 2020).

Appendices

A Ablation Study

This section shows the stability of the DICE model in terms of classification performance by changing different hyper-parameters. This section also shows that as DICE was not fine-tuned extensively for different experiments, it is possible to achieve better classification performance than reported in the relevant chapter. Table A.1 shows the effect of number of test folds on classification performance of DICE. Table A.2 shows the effect on performance when changing the size of hidden dimensions. Also, as FBIRN experiments with 18 fold testing created the biggest leakage, the experiment without leakage was necessary for completeness and shows model performs similarly. All other experiments had leakage of 1-2 subjects whose effect should be insignificant. In Table A.3, we can see that it is possible to get a bit different classification results than ones reported in the main body by permuting the subjects in different order.

Table A.1: The table shows the effect of the different number of cross-validation folds on the classification performance of the DICE model using ICA data. Additional experiment (18, no leakage) where the last fold had all the remaining subjects to prevent any data leakage was also performed. We see that the DICE shows similar performance on different number of cross-validation folds with an increase in performance with a greater number of folds.

Dataset	Number of test folds	Mean AUC	Median AUC
FBIRN	4	0.859	0.861
FBIRN	18	0.86	0.861
FBIRN	18, no leakage	0.86	0.861
ABIDE	5	0.7052	0.71
ABIDE	10	0.722	0.732
OASIS	5	0.741	0.749
OASIS	10	0.752	0.758

Table A.2: The table shows how hidden dimensions of different modules of the model affect classification performance of DICE model. The table shows it is possible to get better results than ones reported in the main body of the paper. Similar results were seen for other datasets as well. We see how removing the temporal attention reduces the model's classification performance. None means the final connectivity matrix $\mathbf{W}^f$ was just the average of each $\mathbf{W}_t$.

Dataset	biLSTM dimension	Self-attention dimension	γ_2	Temporal Attention	Mean AUC	Median AUC
FBIRN	100	48	0.05	GTA	0.86	0.861
FBIRN	100	48	0.05	None	0.733	0.764
FBIRN	100	64	0.05	GTA	0.858	0.861
FBIRN	128	64	0.025	GTA	0.865	0.875
FBIRN	128	64	0.025	None	0.761	0.778
FBIRN	64	32	0.05	GTA	0.849	0.858

Table A.3: Permuting the order of the subjects can lead to a small variation in the classification performance.

Dataset	biLSTM dimension	Self-attention dimension	γ_2	Permutation	Mean AUC	Median AUC
FBIRN	100	48	0.05	Random	0.86	0.861
FBIRN	128	64	0.025	Random	0.865	0.875
FBIRN	100	48	0.05	Default order	0.86	0.889
FBIRN	128	64	0.025	Default order	0.858	0.875

B DNC with negative weights

Connectivity of a node with itself equal to one is the only known and correct bias that can be used while estimating connectivity matrix between nodes. Therefore an additional experiment is done withe the DICE model by adding a new loss term in Equation 5.5 and create following two variations.

$$loss = \text{CrossEntropy}(\hat{\mathbf{y}}, \mathbf{y}) + \beta(1 - \frac{1}{N}\,\text{tr}(\tanh(\mathbf{W}^f))) + \lambda\|\boldsymbol{\theta}\|_1 \tag{B.1}$$

$$loss = \text{CrossEntropy}(\hat{\mathbf{y}}, \mathbf{y}) + \beta(1 - \frac{1}{N}\,\text{tr}(\text{sigmoid}(\mathbf{W}^f))) + \lambda\|\boldsymbol{\theta}\|_1 \tag{B.2}$$

The second term in equations B.1 and B.2 is used to encourage the model to produce connectivity matrices with the average value of the main diagonal closer to 1. `tr` represents the trace of a matrix. β is a regularization coefficient and kept as 0.75. β equal to 1 does push the diagonal closer to 1 but leads to reduction in classification performance. It was found in the experiments that the second term results in more stable and easier to visualize matrices across multiple trials. The added term did not significantly affect the classification performance as shown in Table B.4 with tanh and sigmoid activation. Figure B.1 shows the same matrix as Figure 5.6a created with the new loss equation B.1.

Table B.4: Classification performance of DICE on FBIRN ICA data with the new term added in the loss function. There is not a significant difference in performance, though marginal improvement is seen with sigmoid activation.

Dataset	Added loss term	Mean AUC	Median AUC
FBIRN	None	0.86	0.861
FBIRN	tanh	0.859	0.861
FBIRN	sigmoid	0.862	0.875

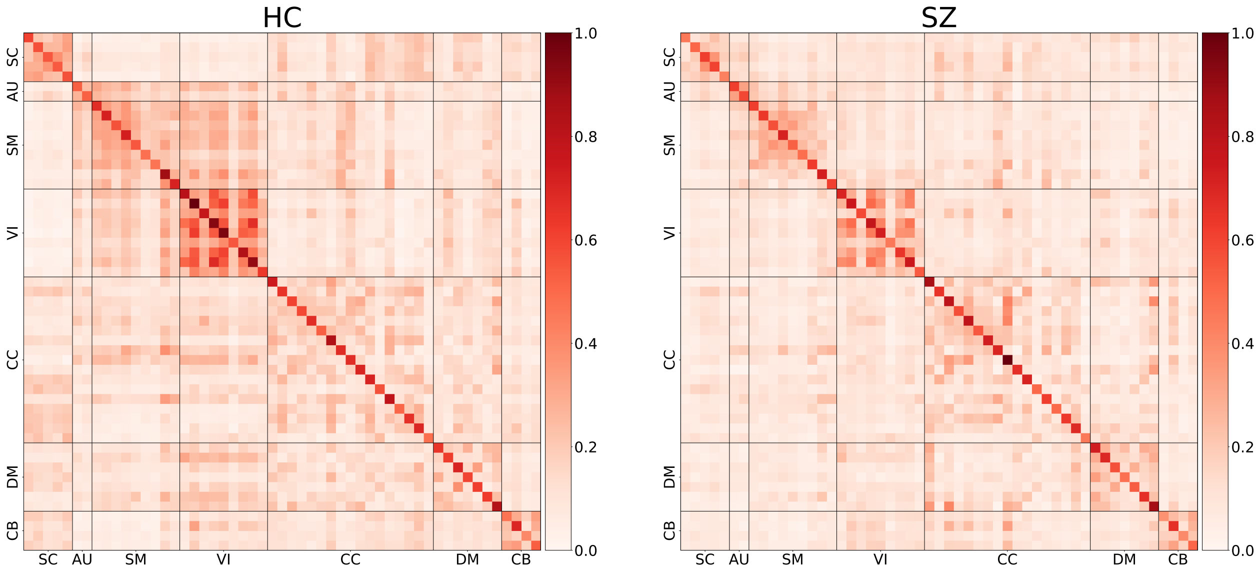
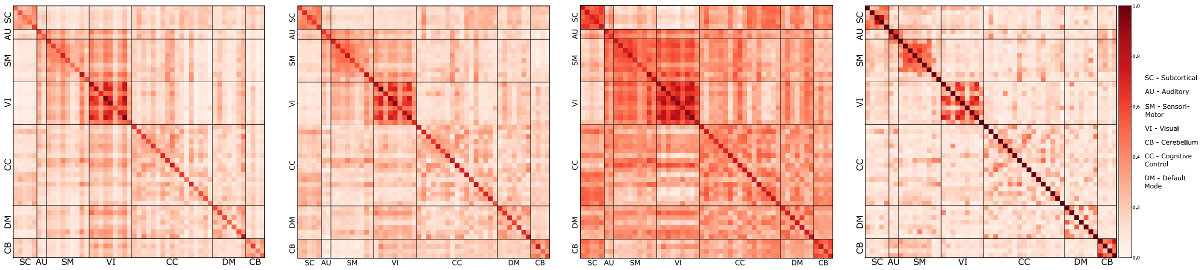


Figure B.1 DNC estimated by DICE model using the loss equation B.1. Same FBIRN subjects as in Figure 5.6a were used for creating this figure.

Figure 5.4 is also re-created using the new loss equations B.1 and B.2 and show the estimated DNC in Figure B.2. The added loss terms noticeably increase the values on the diagonal of the connectivity matrices closer to 1. Notably, the difference between diagonal and non-diagonal values is higher in DNC with tanh loss term than sigmoid based DNC. Presumably, this is probably because the output value for non-negative input (0) in sigmoid is 0.5 and not 0 as in tanh. Hence, the loss for sigmoid is in the range [0-0.5] and not [0-1]. The choice of the function depends on the application and factors such as the presence of self edges, negative edges, the range of the edge weights etc.



(a) DICE DNC

(b) DICE DNC - Tanh

(c) DICE DNC - Sigmoid

(d) PCC FNC

Figure B.2 Comparison of the DNCs learned with the additional regularization terms in the loss function against the DNC created using original loss and PCC FNC. As expected, regularization pushes the diagonal closer to 1. Also the difference between values of diagonal and non-diagonal elements is higher in tanh based DNC B.2b as compared to sigmoid based DNC B.2c. Similarly to Figure 5.4 these matrices are averaged across multiple tries.

As FC and FNC are computed using PCC method to measure the correlations, it has negative correlations as well. These negative correlations are used in different studies and have meaningful interpretations. Therefore, this study also tried to accommodate negative values in the DC and DNC estimated by the DICE model. This can be done easily by making a small tweak in the self-attention part of the model. Equation 5.2 uses softmax function to get the weights and forces them in the range 0-1. Negative weights can be achieved by replacing the softmax function with tanh. Figure 5.4a is re-created by estimating negative weights as well. We see in Figure B.3 that DICE can capture the negative weights by making a small tweak in the self-attention part but detail experiments are required to check the classification performance, stability, and interpretation if negative weights are incorporated. Also, incorporating negative weights require some hyper-parameter changes as well. This is left as future work.

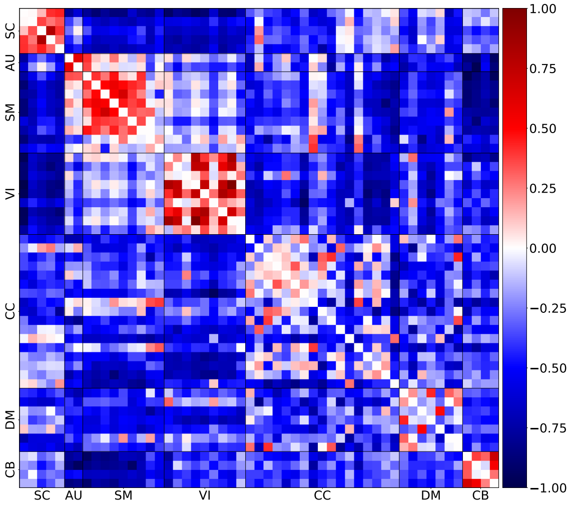


Figure B.3 DNC estimated by DICE model by incorporating negative weights in self-attention module. Same subjects of FBIRN were used as in Figure 5.4a. The diagonal is manually asigned 0 weight.

REFERENCES

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., & Varoquaux, G. 2017, NeuroImage, 147, 736

Aertsen, A., & Preissl, H. 1991, Nonlinear Dynamics and Neuronal Networks, 281

Albers, M. W. et al. 2015a, Alzheimers Dement, 11, 70, [PubMed Central:PMC4287457] [DOI:10.1016/j.jalz.2014.04.514] [PubMed:21768501]

Albers, M. W., Gilmore, G. C., Kaye, J., Murphy, C., et al. 2015b, Alzheimer's & Dementia, 11, 70

Allen, E., Damaraju, E., Plis, S., Erhardt, E., Eichele, T., & Calhoun, V. 2012, Cerebral cortex (New York, N.Y. : 1991)

Allen, E. et al. 2011a, Frontiers in Systems Neuroscience, 5, 2

Allen, E. A. et al. 2011b, Frontiers in systems neuroscience, 5, 2

Anand, A., Racah, E., Ozair, S., Bengio, Y., Côté, M.-A., & Hjelm, R. D. 2019, arXiv preprint arXiv:1906.08226

Angelov, P. P., Soares, E. A., Jiang, R., Arnold, N. I., & Atkinson, P. M. 2021, WIREs Data Mining and Knowledge Discovery, 11, e1424

Armstrong, C. C., Moody, T. D., Feusner, J. D., McCracken, J. T., Chang, S., Levitt, J. G., Piacentini, J. C., & O'Neill, J. 2016, Journal of Affective Disorders, 193, 175

Arslan, S., Ktena, S. I., Glocker, B., & Rueckert, D. 2018, Graph Saliency Maps through Spectral Convolutional Networks: Application to Sex Classification with Brain Connectivity

Bachman, P., Hjelm, R. D., & Buchwalter, W. 2019, Learning Representations by Maximizing Mutual Information Across Views

Bielza, C., & Larranaga, P. 2014, Frontiers in computational neuroscience, 8, 131

Bresson, X., & Laurent, T. 2017, arXiv preprint arXiv:1711.07553

Breukelaar, I. A., Antees, C., Grieve, S. M., Foster, S. L., Gomes, L., Williams, L. M., & Korgaonkar, M. S. 2017, Hum Brain Mapp, 38, 631

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. 2017, IEEE Signal Processing Magazine, 34, 18–42

Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. 2014, Spectral Networks and Locally Connected Networks on Graphs

Butler, P., Silverstein, S., & Dakin, S. 2008, Biological psychiatry, 64, 40

Calhoun, V., Miller, R., Pearlson, G., & Adalı, T. 2014, Neuron, 84, 262

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. 2001, Human brain mapping, 14, 140

Cao, M., Yang, M., Qin, C., Zhu, X., Chen, Y., Wang, J., & Liu, T. 2021, Biomedical Signal Processing and Control, 70, 103015

Casey, B. J. et al. 2018, Developmental cognitive neuroscience, 32, 43

Çetin, M. S. et al. 2014, Neuroimage, 97, 117

Chen, Y., Nakayama, K., Levy, D., Matthysse, S., & Holzman, P. 1999, Proceedings of the National Academy of Sciences, 96

Cheng, J., Dong, L., & Lapata, M. 2016, Long Short-Term Memory-Networks for Machine Reading

Chiang, S., Guindani, M., Yeh, H. J., Haneef, Z., Stern, J. M., & Vannucci, M. 2017, Human brain mapping, 38, 1311

Chickering, D. 2002a, Journal of Machine Learning Research, 3, 507

Chickering, D. M. 2002b, J. Mach. Learn. Res., 2, 445–498

Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y. 2014, in Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (Doha, Qatar: Association for Computational Linguistics), 103–111

Cole, M. W., & Schneider, W. 2007, NeuroImage, 37, 343

Comon, P. 1994, Signal processing, 36, 287

Cordova-Palomera, A. et al. 2017, Scientific Reports, 7

Culbreth, A., Wu, Q., Chen, S., Adhikari, B., Gold, J., & Waltz, J. 2021, NeuroImage: Clinical, 29, 102531

da Xu, chuanwei ruan, evren korpeoglu, sushant kumar, & kannan achan. 2020, in International Conference on Learning Representations

Damaraju, E. et al. 2014, NeuroImage: Clinical, 5, 298

Deshpande, G., Santhanam, P., & Hu, X. 2011, Neuroimage, 54, 1043

Desikan, R. S. et al. 2006, NeuroImage, 31, 968

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2018, arXiv preprint arXiv:1810.04805

Dhurandhar, A., Shanmugam, K., Luss, R., & Olsen, P. 2018, Improving Simple Models with Confidence Profiles

Di Martino, A. et al. 2014, Molecular psychiatry, 19, 659

Douglas, P., Harris, S., Yuille, A., & Cohen, M. S. 2011, NeuroImage, 56, 544

Du, Y., Fu, Z., & Calhoun, V. D. 2018, Frontiers in Neuroscience, 12, 525

Eavani, H., Satterthwaite, T. D., Gur, R. E., Gur, R. C., & Davatzikos, C. 2013, in International Conference on Information Processing in Medical Imaging, Springer, 426–437

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. 2010, Journal of Machine Learning Research, 11, 625

Fedorov, A., Hjelm, R. D., Abrol, A., Fu, Z., Du, Y., Plis, S., & Calhoun, V. D. 2019, arXiv preprint arXiv:1904.10931

Filippi, M. et al. 2017, Neurology, 89, 1764, [PubMed Central:PMC5664301] [DOI:10.1212/WNL.0000000000004577] [PubMed:26888621]

Filippi, M., Valsasina, P., Misci, P., Falini, A., Comi, G., & Rocca, M. 2013, Human brain mapping, 34

Freedman, D., Pisani, R., & Purves, R. 2007, Pisani, R. Purves, 4th edn. WW Norton & Company, New York

Friston, K. 2011, Brain connectivity, 1, 13

Fu, Z., Caprihan, A., Chen, J., Du, Y., Adair, J. C., Sui, J., Rosenberg, G. A., & Calhoun, V. D. 2019, Human Brain Mapping

Fu, Z., Iraji, A., Turner, J. A., Sui, J., Miller, R., Pearlson, G. D., & Calhoun, V. D. 2021a, NeuroImage, 224, 117385

Fu, Z., Sui, J., Turner, J., Du, Y., Assaf, M., Pearlson, G., & Calhoun, V. 2020, Human brain mapping, 42

Fu, Z., Sui, J., Turner, J. A., Du, Y., Assaf, M., Pearlson, G. D., & Calhoun, V. D. 2021b, Human Brain Mapping, 42, 80

Fu, Z. et al. 2018, NeuroImage, 190

Gadgil, S., Zhao, Q., Pfefferbaum, A., Sullivan, E. V., Adeli, E., & Pohl, K. M. 2021, Spatio-Temporal Graph Convolution for Resting-State fMRI Analysis

Gao, H., & Ji, S. 2019, Graph U-Nets

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. 2020, Nature Machine Intelligence, 2, 665

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. 2017, Neural Message Passing for Quantum Chemistry

Glasser, M. et al. 2013, NeuroImage, 80, 105

Goebel, R., Roebroeck, A., Kim, D.-S., & Formisano, E. 2003, Magnetic resonance imaging, 21, 1251

Gorrostieta, C., Fiecas, M., Ombao, H., Burke, E., & Cramer, S. 2013, Frontiers in computational neuroscience, 7, 159

Grant, A., Dennis, N. A., & Li, P. 2014, Front Psychol, 5, 1401, [PubMed Central:PMC4253532] [DOI:10.3389/fpsyg.2014.01401] [PubMed:15322270]

Griffanti, L. et al. 2014, NeuroImage, 95, 232

Guo, W., Liu, F., Chen, J., Wu, R., Li, L., Zhang, Z., Chen, H., & Zhao, J. 2017, Medicine, 96, e6223

Haan, W., Flier, W., Koene, T., Smits, L., Scheltens, P., & Stam, C. 2011, NeuroImage, 59, 3085

Hamilton, W. L., Ying, R., & Leskovec, J. 2018, Representation Learning on Graphs: Methods and Applications

Hasani, R., Lechner, M., Amini, A., Rus, D., & Grosu, R. 2020, Liquid Time-constant Networks

Hénaff, O. J., Razavi, A., Doersch, C., Eslami, S., & Oord, A. v. d. 2019, arXiv preprint arXiv:1905.09272

Hjelm, R. D., Calhoun, V. D., Salakhutdinov, R., Allen, E. A., Adali, T., & Plis, S. M. 2014, NeuroImage, 96, 245

Hjelm, R. D., Damaraju, E., Cho, K., Laufs, H., Plis, S. M., & Calhoun, V. D. 2018a, Frontiers in neuroscience, 12, 600

Hjelm, R. D., Fedorov, A., Lavoie-Marchildon, S., Grewal, K., Bachman, P., Trischler, A., & Bengio, Y. 2018b, arXiv preprint arXiv:1808.06670

Hutchison, R., Womelsdorf, T., Gati, S., Everling, S., & Menon, R. 2013, Human brain mapping, 34

Hyvärinen, A., & Oja, E. 2000, Neural Networks, 13, 411

Ingalhalikar, M. et al. 2014, Proceedings of the National Academy of Sciences, 111, 823

Jacobs, H. I. L., Hopkins, D. A., Mayrhofer, H. C., Bruner, E., van Leeuwen, F. W., Raaijmakers, W., & Schmahmann, J. D. 2017, Brain, 141, 37

Jafri, M. J., Pearlson, G. D., Stevens, M., & Calhoun, V. D. 2008, Neuroimage, 39, 1666

Jain, S., & Wallace, B. C. 2019, Attention is not Explanation

Jalil, K. A., Kamarudin, M. H., & Masrek, M. N. 2010, in 2010 International Conference on Networking and Information Technology, 221–226

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. 2012, NeuroImage, 62, 782, 20 YEARS OF fMRI

Johansson, U., Sönströd, C., Norinder, U., & Boström, H. 2011, Future Medicinal Chemistry, 3, 647, pMID: 21554073

Just, M. A., Keller, T. A., Malave, V. L., Kana, R. K., & Varma, S. 2012, Neuroscience & Biobehavioral Reviews, 36, 1292

Kan, X., Dai, W., Cui, H., Zhang, Z., et al. 2022, ArXiv, abs/2210.06681

Kawahara, J., Brown, C., Miller, S., Booth, B., Chau, V., Grunau, R., Zwicker, J., & Hamarneh, G. 2016, NeuroImage, 146

Kazi, A., Farghadani, S., & Navab, N. 2021, IA-GCN: Interpretable Attention based Graph Convolutional Network for Disease prediction

Keator, D. B. et al. 2016, Neuroimage, 124, 1074

Khosla, M., Jamison, K., Kuceyeski, A., & Sabuncu, M. R. 2019a, in International Workshop on Machine Learning in Medical Imaging, Springer, 301–309

Khosla, M., Jamison, K., Ngo, G. H., Kuceyeski, A., & Sabuncu, M. R. 2019b, Magnetic resonance imaging

Kim, B.-H., & Ye, J. C. 2020, Frontiers in Neuroscience, 14, 630

Kim, B.-H., Ye, J. C., & Kim, J.-J. 2021, Learning Dynamic Graph Representation of Brain Connectome with Spatio-Temporal Attention

Kipf, T., Fetaya, E., Wang, K.-C., Welling, M., & Zemel, R. 2018, Neural Relational Inference for Interacting Systems

Kipf, T. N., & Welling, M. 2017, Semi-Supervised Classification with Graph Convolutional Networks

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. 2021, Mechanical Systems and Signal Processing, 151, 107398

Knyazev, B., Taylor, G. W., & Amer, M. R. 2019, Understanding Attention and Generalization in Graph Neural Networks

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., & Rueckert, D. 2017, Distance Metric Learning using Graph Convolutional Networks: Application to Functional Brain Networks

——. 2018, NeuroImage, 169, 431

Kéri, S., Antal, A., Szekeres, G., Benedek, G., & Janka, Z. 2002, The Journal of Neuropsychiatry and Clinical Neurosciences, 14, 190, pMID: 11983794

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proceedings of the IEEE, 86, 2278

Lewis, N., Miller, R., Gazula, H., Rahman, M. M., Iraji, A., Calhoun, V. D., & Plis, S. 2021, in 2021 IEEE 9th International Conference on Healthcare Informatics (ICHI), 243–247

Li, H., Parikh, N. A., & He, L. 2018, Frontiers in Neuroscience, 12, 491

Li, Y., Zemel, R., Brockschmidt, M., & Tarlow, D. 2016, in Proceedings of ICLR'16, proceedings of iclr'16 edn.

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. 2017, A Structured Self-attentive Sentence Embedding

Liu, Z., Adeli, E., Pohl, K., & Zhao, Q. 2021

Lugosch, L., Ravanelli, M., Ignoto, P., Tomar, V. S., & Bengio, Y. 2019, arXiv preprint arXiv:1904.03670

Luo, Y., Tseng, H.-H., Cui, S., Wei, L., Ten Haken, R. K., & El Naqa, I. 2019, BJR—Open, 1, 20190021

Lütkepohl, H. 2005, New introduction to multiple time series analysis (Springer Science & Business Media)

Lynall, M.-E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., & Bullmore, E. 2010, Journal of Neuroscience, 30, 9477

Ma, G., Ahmed, N. K., Willke, T., Sengupta, D., Cole, M. W., Turk-Browne, N. B., & Yu, P. S. 2019, Similarity Learning with Higher-Order Graph Convolutions for Brain Network Analysis

Mahmood, U., Fu, Z., Calhoun, V., & Plis, S. 2021a, Multi network InfoMax: A pre-training method involving graph convolutional networks

——. 2022, Deep Dynamic Effective Connectivity Estimation from Multivariate Time Series

Mahmood, U., Fu, Z., Calhoun, V. D., & Plis, S. 2021b, Algorithms, 14, 75

Mahmood, U., Rahman, M. M., Fedorov, A., Fu, Z., Calhoun, V. D., & Plis, S. M. 2019a, Learnt dynamics generalizes across tasks, datasets, and populations

Mahmood, U., Rahman, M. M., Fedorov, A., Fu, Z., & Plis, S. 2019b, Transfer Learning of fMRI Dynamics, machine Learning for Health (ML4H) at NeurIPS 2019 - Extended Abstract

Mahmood, U., Rahman, M. M., Fedorov, A., Lewis, N., Fu, Z., Calhoun, V. D., & Plis, S. M. 2020a, Lecture Notes in Computer Science, 407–417

Mahmood, U., Rahman, M. M., Fedorov, A., Lewis, N., Fu, Z., Calhoun, V. D., & Plis, S. M. 2020b, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, ed. A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, & L. Joskowicz (Cham: Springer International Publishing), 407–417

Mak, L., Minuzzi, L., MacQueen, G., Hall, G., Kennedy, S., & Milev, R. 2016, Brain Connectivity, 7

Marutho, D., Hendra Handaka, S., Wijaya, E., & Muljono. 2018, in 2018 International Seminar on Application for Technology of Information and Communication, 533–538

Mensch, A., Mairal, J., Bzdok, D., Thirion, B., & Varoquaux, G. 2017, in Advances in Neural Information Processing Systems, 5883–5893

Miller, R., & Calhoun, V. 2020a, in Medical Imaging 2020, ed. I. Isgum & B. Landman, Progress in Biomedical Optics and Imaging - Proceedings of SPIE (SPIE), publisher Copyright: © 2020 SPIE. All rights reserved. Copyright: Copyright 2020 Elsevier B.V., All rights reserved.; Medical Imaging 2020: Image Processing ; Conference date: 17-02-2020 Through 20-02-2020

Miller, R. L., & Calhoun, V. D. 2020b, in 2020 IEEE Southwest Symposium on Image

Analysis and Interpretation (SSIAI), 108–111

Million, E. 2007, The Hadamard Product

Mitra, A., Snyder, A. Z., Hacker, C. D., & Raichle, M. E. 2014, Journal of Neurophysiology, 111, 2374, pMID: 24598530

Monti, F., Boscaini, D., Masci, J., Rodolà, E., Svoboda, J., & Bronstein, M. M. 2016, Geometric deep learning on graphs and manifolds using mixture model CNNs

Morgan, S. et al. 2020a, Biological Psychiatry: Cognitive Neuroscience and Neuroimaging

——. 2020b, Biological Psychiatry: Cognitive Neuroscience and Neuroimaging

Oord, A. v. d., Li, Y., & Vinyals, O. 2018, arXiv preprint arXiv:1807.03748

Parikh, A. P., Täckström, O., Das, D., & Uszkoreit, J. 2016, A Decomposable Attention Model for Natural Language Inference

Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., & Rueckert, D. 2018, Medical Image Analysis, 48, 117

Paulus, R., Xiong, C., & Socher, R. 2017, A Deep Reinforced Model for Abstractive Summarization

Pearl, J. 2000, Causality, by Judea Pearl, pp. 400. ISBN 0521773628. Cambridge, UK: Cambridge University Press, March 2000., -1

Plis, S. M. et al. 2014, Frontiers in Neuroscience, 8

Rabany, L. et al. 2019, NeuroImage: Clinical, 24, funding Information: This work has been supported by the National Institutes of Health (NIMH; R01 MH095888 ; PI: M. Assaf), and the National Alliance for Research in Schizophrenia and Affective Disorders (NARSAD;

Young Investigator Award 17525; PI: C. Corbera). Publisher Copyright: © 2019

Ras, G., Xie, N., van Gerven, M., & Doran, D. 2021, Explainable Deep Learning: A Field Guide for the Uninitiated

Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., & Calhoun, V. D. 2016, NeuroImage, 134, 645

Rashid, B., Damaraju, E., Pearlson, G., & Calhoun, V. 2014, Frontiers in human neuroscience, 8, 897

Ravanelli, M., & Bengio, Y. 2018, arXiv preprint arXiv:1812.00271

Ritchie, S. J. et al. 2018, Cerebral Cortex, 28, 2959

Rubin, E. H., Storandt, M., Miller, J. P., Kinscherf, D. A., Grant, E. A., Morris, J. C., & Berg, L. 1998, Archives of neurology, 55, 395

Saha, D. K., Damaraju, E., Rashid, B., Abrol, A., Plis, S. M., & Calhoun, V. D. 2020, bioRxiv

Sakoğlu, Ü., Pearlson, G., Kiehl, K., Wang, Y., Michael, A., & Calhoun, V. 2010, Magnetic Resonance Materials in Physics, Biology, and Medicine, 23, 351, funding Information: Acknowledgments This work was funded by National Institution of Health (NIH)/National Institute of Biomedical Imaging and Bio Engineering (NIBIB) grant 2RO1 EB000840-06 (Calhoun) and National Institute of Mental Health (NIMH) grant RO1 MH072681 (Kiehl). The authors would like to thank the Medical Image Analysis Lab staff (http://mialab.mrn.org) at Mind Research Network (MRN) for their valuable feedback.

Salehi, M., Greene, A. S., Karbasi, A., Shen, X., Scheinost, D., & Constable, R. T. 2020, NeuroImage, 208, 116366

Salimi-Khorshidi, G., Douaud, G., Beckmann, C., Glasser, M., Griffanti, L., & Smith, S. 2014, NeuroImage, 90

Salman, M. S. et al. 2019, NeuroImage: Clinical, 22, 101747

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. 2009, IEEE Transactions on Neural Networks, 20, 61

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. T. 2017, Cerebral Cortex, 28, 3095

Schreiber, T. 2000, Phys. Rev. Lett., 85, 461

Schuster, M., & Paliwal, K. 1997, IEEE Transactions on Signal Processing, 45, 2673

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. 2021, Towards Causal Representation Learning

Seth, A., Barrett, A., & Barnett, L. 2015, The Journal of neuroscience : the official journal of the Society for Neuroscience, 35, 3293

Shukla, P., & Tripathi, S. 2012, Information, 3, 256

Silverstein, S. M., & Rosen, R. 2015, Schizophrenia Research: Cognition, 2, 46, visual Functioning and Schizophrenia

Simonyan, K., Vedaldi, A., & Zisserman, A. 2014, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps

Spirtes, P., & Glymour, C. 1991, Social Science Computer Review - SOC SCI COMPUT REV, 9, 62

Spirtes, P., Glymour, C., & Scheines, R. 1993, Causation, Prediction, and Search, Vol. 81

Supekar, K., Cai, W., Krishnadas, R., Palaniyappan, L., & Menon, V. 2019, Biological Psychiatry, 85, 60, immune Mechanisms and Psychosis

Thomas, A. W., Müller, K.-R., & Samek, W. 2019, arXiv preprint arXiv:1907.01953

Tsai, C.-F. et al. 2019, Scientific Reports, 9

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. 2002, NeuroImage, 15, 273

Ulloa, A., Plis, S., & Calhoun, V. 2018, arXiv preprint arXiv:1804.04591

Ursino, M., Ricci, G., & Magosso, E. 2020, Frontiers in Computational Neuroscience, 14

van den Heuvel, M. P., Mandl, R. C. W., Stam, C. J., Kahn, R. S., & Hulshoff Pol, H. E. 2010, Journal of Neuroscience, 30, 15915

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. 2013, Neuroimage, 80, 62

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., & Polosukhin, I. 2017, in Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17 (Red Hook, NY, USA: Curran Associates Inc.), 6000–6010

Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. 2018, International Conference on Learning Representations

Vicente, R., Wibral, M., Lindner, M., & Pipa, G. 2011, Journal of computational neuroscience, 30, 45

Vinyals, O., Bengio, S., & Kudlur, M. 2016, Order Matters: Sequence to sequence for sets

Wang, X. et al. 2014, Schizophrenia Research, 156, 150

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. 2019, Dynamic Graph CNN for Learning on Point Clouds

Weis, S., Patil, K. R., Hoffstaedter, F., Nostro, A., Yeo, B. T. T., & Eickhoff, S. B. 2019, Cerebral Cortex, 30, 824

Wiegreffe, S., & Pinter, Y. 2019, Attention is not not Explanation

Williams, N., Zander, S., & Armitage, G. 2006, Computer Communication Review, 36, 5

Yaesoubi, M., Adalı, T., & Calhoun, V. D. 2018, Human Brain Mapping, 39, 1626

Yahata, N. et al. 2016, Nature communications, 7, 1

Yan, W., Plis, S., Calhoun, V. D., Liu, S., Jiang, R., Jiang, T.-Z., & Sui, J. 2017, in 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 1–6

Yang, W., Xu, X., Wang, C., Cheng, Y., Li, Y., Xu, S., & Li, J. 2022, Brain Imaging and Behavior

Yao, D., Sui, J., Yang, E., Yap, P.-T., Shen, D., & Liu, M. 2020, in Machine Learning in Medical Imaging, ed. M. Liu, P. Yan, C. Lian, & X. Cao (Cham: Springer International Publishing), 1–10

Yu, Q., Sui, J., Rachakonda, S., He, H., Pearlson, G., & Calhoun, V. 2011, Frontiers in systems neuroscience, 5, 7

Zeng, K. et al. 2017, Scientific Reports, 7

Zhang, C., Dougherty, C., Baum, S., White, T., & Michael, A. 2018a, Human Brain Mapping, 39

Zhang, J. et al. 2018b, Frontiers in Behavioral Neuroscience, 12

Zhang, M., & Chen, Y. 2018, Link Prediction Based on Graph Neural Networks

Zhang, Y., Dai, Z., Chen, Y., Sim, K., Sun, Y., & Yu, R. 2019, Brain Imaging and Behavior, 13

Zhu, J., Qian, Y., Zhang, B., Li, X., Bai, Y., Li, X., & Yu, Y. 2020, Brain Imaging and Behavior, 14

Zitnik, M., Agrawal, M., & Leskovec, J. 2018, Bioinformatics, 34, i457–i466