



Learning Multi-resolution Graph Edge Embedding for Discovering Brain Network Dysfunction in Neurological Disorders

Xin Ma^{1(✉)}, Guorong Wu², Seong Jae Hwang³, and Won Hwa Kim^{1,4}

¹ University of Texas at Arlington, Arlington, USA

xin.ma@mavs.uta.edu

² University of North Carolina at Chapel Hill, Chapel Hill, USA

³ University of Pittsburgh, Pittsburgh, USA

⁴ POSTECH, Pohang-si, South Korea

Abstract. Tremendous recent literature show that associations between different brain regions, i.e., brain connectivity, provide early symptoms of neurological disorders. Despite significant efforts made for graph neural network (GNN) techniques, their focus on graph nodes makes the state-of-the-art GNN methods not suitable for classifying brain connectivity as graphs where the objective is to characterize disease-relevant network dysfunction patterns on graph links. To address this issue, we propose Multi-resolution Edge Network (MENET) to detect disease-specific connectomic benchmarks with high discrimination power across diagnostic categories. The core of MENET is a novel graph edge-wise transform that we propose, which allows us to capture multi-resolution “connectomic” features. Using a rich set of the connectomic features, we devise a graph learning framework to jointly select discriminative edges and assign diagnostic labels for graphs. Experiments on two real datasets show that MENET accurately predicts diagnostic labels and identify brain connectivities highly associated with neurological disorders such as Alzheimer’s Disease and Attention-Deficit/Hyperactivity Disorder.

1 Introduction

Many neuroimaging studies operate with data from a population of cohort that can be stratified into two or more groups (e.g., diseased vs. control). Given registered imaging measures acquired from participants, contrasting the different groups at each pixel/voxel over the whole brain identifies those regions that are affected by the variable of interest (e.g., disease or risk factors) [8]. However, due to poor correlation between cognitive changes and pathological features from images, recent studies motivate that characterizing changes in the *brain connectivity* or *network* that comprise several affected regions yield a better understanding of the brain over traditional spatial analyses [3, 11, 21].

In brain connectivity studies, an individual brain is divided into registered regions of interests (ROIs) and associations across the ROIs are defined via

functional/structural images, formulated as a “graph” that consists of nodes and edges. Changes in individual connectivity have shown strong implication of neurological diseases with evidence such as Parkinson’s [27] and Alzheimer’s disease (AD) [6]. Still, few frameworks exist for traditional prediction tasks for registered but topologically variant graphs *without node-specific features*.

Existing techniques including Graph Neural Network (GNN) methods mainly focus on analyzing features defined on graph nodes, where the graph is given as the domain of the signal [18, 34]. GNNs for point clouds may work but their nodes are merely arbitrarily sampled points over surfaces of non-rigid shapes [19]. Notice that such settings are not in the interest of this study where we deal with alternations in the “connectivity”, i.e., edges, instead of measures at each node. Graph kernels that compare local substructures [28, 31] are often inadequate for training, and alternative deep learning methods for graph such as [32, 37] lose interpretability in space particularly with dense graphs. For brain connectivity analysis, the method must be able to investigate local edge-wise variation and predict global diagnostic labels to identify disease-specific symptoms. Hence, it is critical to develop a framework that can *adaptively* transform graph “edges” with graphs’ inherent *structure* of the graphs to increase *sensitivity* but maintain location-wise *interpretability* even with small sample sizes.

To achieve the objectives above, we propose a new convolution neural network (CNN) framework for graph edges that derives flexible multi-resolution features sensitive to topological variations in graphs (example shown in Fig. 1). The key of our method is at the adaptive graph edge transform—adopting ideas from spectral graph wavelet transform [12], we define a novel *multi-resolution edge trans-*

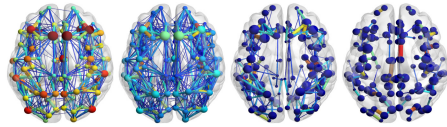


Fig. 1. Example of multi-resolution representation of brain connectivity. 1) original network, 2)-4) filtered network at $s = 0.3, 0.8, 1.8$. Sparsity of the network increases and nodal degree decreases (red to blue) as the scale changes. (Color figure online)

form that deals with a positive semi-definite (p.s.d.) matrix. Defining a graph specific orthonormal tensor as a surrogate to transform a p.s.d. matrix, we obtain its multi-resolution representation with trainable kernel functions in a dual space. Such an *edge-wise* analysis is at the core of brain connectivity analysis identifying which of the brain connectivities are significantly related to disease-specific variables. The **contributions** of our work are, 1) defining a novel graph edge transform that derives flexible representation of graph edges, 2) proposing a framework that can efficiently train and classify diagnostic labels for graphs as well as identify disease-specific brain connectivity, 3) extensive empirical results on two independent real brain connectivity data for AD and Attention-Deficit/Hyperactivity Disorder (ADHD) to validate our framework. Our discoveries identifying subtle variations in brain connectivity align with other on-going studies, suggesting that our framework has potential for various brain disorder analyses.

1.1 Related Work

Graph Kernels (GK): GK methods map graphs into a Hilbert space, which can be fed to downstream classifiers (e.g., Support Vector Machine (SVM)) to perform graph classification [10]. The designs of kernels include neighborhood aggregation (e.g., Weisfeiler-Lehman [30]), extraction of subgraph patterns (e.g., Graphlet [31], and walks/paths (e.g., shortest-path [1] and random walk [14]).

GNN: Graph Convolution Network (GCN) methods transform “node features” into informative node embeddings by aggregating and propagating node features via graph convolution. Such node embeddings have shown to excel at node classification [18] and graph classifications [34] with specific pooling mechanisms, e.g., Dynamic Graph CNN (DGCNN), one of the state-of-the-art methods, uses a specialized SortPooling layer for graph classification [35]. Despite their success, GCNs and their variants rely heavily on the node features provided by the data which, in fact, may not even be provided in certain datasets. In other words, graphs may *only* come with edge features, leading to sub-optimal use of GCN by reluctantly utilizing weak node features such as nodal degrees.

Benefits of Our Work. Unlike the two approaches above, we use parametric kernels can adaptively derive beneficial representations for graph classification that can be directly applied to edge measures instead of node features. Different from the methods that define multi-resolution in the native graph space with hops on nodes [36] or thresholds on edges [13], we use a dual space and smooth kernels that can be described by the theory of traditional wavelet transform.

2 Proposed Method

The key in our approach is to derive the multi-resolution representation of *edges* by developing a novel graph edge transform, which we introduce below.

2.1 Multi-resolution Graph Edge Transform

Consider a set of graphs with registered nodes with label y assigned for each graph G . Each graph with N nodes is represented as its adjacency matrix $A_{N \times N}$, whose non-negative element a_{ij} denotes edge weight (i.e., measure of association) between the i -th and j -th nodes. Given an adjacency matrix A (e.g., a brain connectivity), one can derive a degree matrix $D_{N \times N}$ as a diagonal matrix whose i -th diagonal is the sum of edge weights connected to the i -th node (i.e., volume). Its graph Laplacian \mathcal{L} is defined as $D - A$, and a normalized one is defined as $\mathcal{L}_{norm} = D^{-1/2} \mathcal{L} D^{-1/2}$. Both Laplacians, \mathcal{L} and \mathcal{L}_{norm} , are symmetric and p.s.d. with non-negative eigenvalues λ_ℓ and orthonormal eigenvectors u_ℓ . Previous works used the λ_ℓ and u_ℓ to define a wavelet transform on node signals and constructed GCNs for signals defined on the nodes n , i.e., $f(n)$ [12, 18]. However, notice that $f(n)$ is not of our interest but we define a transform for $f(e)$, i.e., weights on the edges e , to derive “multi-resolution” views.

Let us first decompose a graph Laplacian $\mathcal{L} = \sum_{l=0}^{N-1} \lambda_l u_l u_l^T = U \Lambda U^T$ where $U = [u_0, u_1, \dots, u_{N-1}]$ and Λ is a diagonal matrix whose $(l+1)$ -th diagonal is λ_l . Traditional continuous wavelet transform suggests that derive multi-resolution representation of signals can be derived using an orthogonal transform and defining scales in a dual space [12, 17, 22]. Hence, we define an orthonormal basis for a “matrix” using an outer product of u_l , together with a kernel $k(\cdot)$ as $\psi_{l,s}(i, j) = k(s\lambda_l)u_l(i)u_l(j)$ and a transform using the $\psi_{l,s}(i, j)$ is defined as

$$\beta_{\mathcal{L},l}(s) = \langle \mathcal{L}, \psi_l \rangle = \sum_i \sum_j k(s\lambda_l)u_l(i)u_l(j) \sum_{l'=0}^{N-1} \lambda_{l'} u_{l'}(i)u_{l'}(j) = \lambda_l k(s\lambda_l) \quad (1)$$

which yields a resultant coefficient $\beta_{\mathcal{L},\ell}(s)$. Note that its inverse is

$$\mathcal{L}(i, j) = \frac{1}{C_k} \int_0^\infty \sum_{l=0}^{N-1} \beta_{\mathcal{L},l}(s) \psi_{s,l}(i, j) \frac{ds}{s} \quad (2)$$

which reconstructs \mathcal{L} with a kernel normalization constant $C_k = \int_0^\infty \frac{k(x)^2}{x} dx$.

Lemma 1 (*Graph Laplacian Admissibility Condition*). *Given a kernel dependent normalization constant $C_k = \int_0^\infty \frac{k(x)^2}{x} dx < \infty$, the original graph Laplacian \mathcal{L} can be perfectly reconstructed via the inverse transformation.*

Proof. Projecting the coefficients $\beta_{\mathcal{L},l}(s)$ back to the original domain with $\psi_{l,s}$,

$$\begin{aligned} \frac{1}{C_k} \int_0^\infty \sum_{l=0}^{N-1} \beta_{\mathcal{L},l}(s) \psi_{l,s}(i, j) \frac{ds}{s} &= \frac{1}{C_k} \int_0^\infty \sum_{l=0}^{N-1} \lambda_l k(s\lambda_l)^2 u_l(i)u_l(j) \frac{ds}{s} \\ &= \sum_{l=0}^{N-1} \lambda_l \left[\frac{1}{C_k} \int_0^\infty \frac{k(s\lambda_l)^2}{s} ds \right] u_l(i)u_l(j) = \mathcal{L}(i, j) \end{aligned}$$

and $\mathcal{L}(i, j)$ is obtained by substituting $s\lambda_l = x$.

The Lemma above follows the traditional admissibility condition in wavelet transform [22], i.e., a superposition of multi-resolution representation of \mathcal{L} over scales s . It lets us to define new representation \mathcal{L}_s at different scales \mathbf{s} :

$$\mathcal{L}_s(i, j) = \sum_{l=0}^{N-1} \lambda_l k(s\lambda_l)^2 u_l(i)u_l(j) \quad (3)$$

by focusing at a specific scale s . It is given as a matrix operation as

$$\mathcal{L}_s = U k(s\Lambda)^2 \Lambda U^T \quad (4)$$

where $k(s\Lambda)^2$ is a diagonal matrix with the $k(\cdot)$ applied at each λ . The shape of $k(\cdot)$ determines the shape of wavelet-like basis and multi-resolution views of \mathcal{L} . For regular images, this yields filtered images (e.g., band-pass filtered signal), and our framework extracts similar representations of edges $f(e)$ as in Fig. 1.

2.2 Efficient Graph Matrix Transform

The transform (3) requires eigendecomposition of a graph Laplacian (or estimating a partial set of eigenvectors) which can be computationally burdening, especially with a large number of nodes or graph samples. We therefore suggest an approximation of (3) that significantly reduces computation with marginal error, which extends the approximation in [12] to our graph matrix transform.

For this, let us assume $g(\Lambda) = \Lambda k(s\Lambda)^2$ from (4) and $\tilde{\Lambda} = \frac{2}{\lambda_{max}}\Lambda - I_N$ since the largest eigenvalue of a normalized graph Laplacian is bounded by 2. Then, $\tilde{\Lambda} = \Lambda - I_N$ and $g(\tilde{\Lambda}) = g(\Lambda - I_N)$. If we expand $g(\Lambda)$ with I_N then,

$$g(\Lambda) = \sum_{n=0}^{\infty} \frac{g^{(n)}(I_N)}{n!} (\Lambda - I_N)^n = \sum_{n=0}^{\infty} \frac{g^{(n)}(0_N)}{n!} \tilde{\Lambda}^n. \quad (5)$$

Considering that the elements of $\tilde{\Lambda}$ are in $[-1, 1]$, we assume there exists a positive K where if $n > K$, $\tilde{\Lambda}^n \rightarrow 0$. Here, $g(\Lambda)$ can be approximated as

$$g(\Lambda) \approx \sum_{n=0}^K \frac{g^{(n)}(0_N)}{n!} \tilde{\Lambda}^n. \quad (6)$$

Since $U\Lambda^m U^T = (U\Lambda U^T)^m$, \mathcal{L}_s now can be written as

$$\mathcal{L}_s = U g(\Lambda) U^T \approx \sum_{n=0}^K \frac{g^{(n)}(0_N)}{n!} U \tilde{\Lambda}^n U^T \approx \sum_{n=0}^K \frac{g^{(n)}(0_N)}{n!} \tilde{\mathcal{L}}^n \quad (7)$$

where $\tilde{\mathcal{L}} = \mathcal{L} - I_{N \times N}$. This approximation lets our overall framework more practical to handle large scale graph data.

2.3 Network Architecture

Based on the matrix transform (3) as a convolution, we propose *Multi-resolution Edge Network (MENET)* which is a novel CNN framework that utilizes multi-resolution representations of graph edges for classification. Figure 2 illustrates the overall pipeline with the following components:

Convolution Layer: Taking a graph as an input, it returns multi-resolution representation \mathcal{L}_s of its graph Laplacian using the transform in (4) as a tensor, i.e., $|\mathbf{s}| \times N \times N$ where $|\mathbf{s}|$ is the total number of scales. It maps \mathcal{L} to a high-dimensional space of filtered graph edges with $|\mathbf{s}|$.

FC Layer: This component is a fully connected Deep Neural Network (DNN) classifier that takes the multi-resolution features of a graph as an input and predicts its class label. It has L output units where L is the number of available classes. The values o_l computed at each output unit, when normalized, become the pseudo-probability of an input belong to a specific class.

MENET learns two sets of parameters: 1) scale \mathbf{s} that define the resolutions of graph edges, and 2) weight \mathbf{W}^h within the FC for prediction. Notice that there

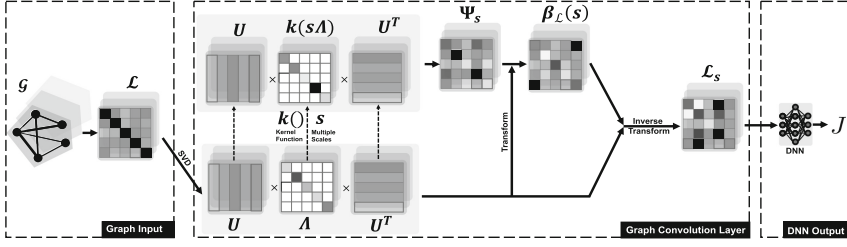


Fig. 2. Overall architecture of MENET. A graph matrix is transformed to yield multi-resolution representations, and then fully connected (FC) DNN is applied at the end. Error is backpropagated to train the weights \mathbf{W}^h and update the scales \mathbf{s} to obtain the optimal representations.

is no pooling; while it is important to increase the efficiency of the algorithm, without an invertible method for pooling on edges, we want to keep *spatial interpretability* for neuroimaging applications. Also, the pooling in conventional methods are required for multi-scale analysis with fixed window size, but our framework inherits such behavior within the matrix transform.

2.4 Training MENET

Given a training set with N_G number of individual graphs G with corresponding labels y , the learning process consists of feedforward and backpropagation steps. In the feedforward propagation of MENET, a graph G is inputted to the framework as a graph Laplacian, and the probability that the G belongs to a particular class is computed at the output. Suppose we are given an input graph G with N vertices and a set of initial scales \mathbf{s} . First, using the operation in (4), the input G is transformed to \mathcal{L}_s . Since each \mathcal{L}_s is represented as a matrix and there are $|\mathbf{s}|$ of them, \mathcal{L}_s are combined to consist a feature map M as a tensor:

$$M = \mathcal{L}_{s_0} \cup \mathcal{L}_{s_1} \cup \dots \cup \mathcal{L}_{s_{|\mathbf{s}|}} \quad (8)$$

Given T hidden units in the hidden layers of the DNN module, at each t -th hidden unit, the learned features in hidden units are linearly combined with corresponding weights in the first layer of DNN as

$$z_t = \sum_{s,p,q}^{|\mathbf{s}|,N,N} w_{t,spq}^h m_{spq}, \quad (9)$$

where m_{spq} represents a feature element in M (i.e., an element at (p, q) in s -th scale) and $w_{t,spq}^h$ denotes the weight on a connection between a hidden unit m_{spq} to the t -th hidden unit of the following layer. A non-linear activation function $\sigma()$ (e.g., sigmoid or rectified linear function) is applied on z_t and is then fed to the next layer. At the output of the DNN, a soft-max function is used to get the

final outcome o_l for the l -th output unit as a pseudo probability from which a prediction is made as $\hat{y} = \arg \max_l o_l$.

Once the output o_l is obtained from the feedforward system above, an error can be computed between the o_l and the target value y_l , i.e., a label with one-hot-encoding for the input graph. We use focal loss [20] to measure the error:

$$J(\mathbf{s}, \mathbf{W}^h) = -\frac{1}{N_G} \sum_{i=1}^{N_G} \sum_{l=1}^L -\alpha_{il}(1 - p_{il})^\gamma \log(p_{il}) \quad (10)$$

where $p_{il} = o_{il}$ if target value $y_{il} = 1$ at l -th output unit with i -th sample, otherwise $p_{il} = 1 - o_{il}$, N_G and L are the total number of graphs in a batch and the total number of available classes respectively, α and γ are balanced variants. Our framework ‘‘adaptively’’ learns scale parameters \mathbf{s} for novel graph representations and \mathbf{W}^h in the FC layer (embedded in o_l) by primarily minimizing the classification error in (10) via backpropagation. In traditional wavelet transform, the \mathbf{s} is fixed to yield theoretical guarantees, however, we freely explore different aspects of \mathbf{s} to find the optimal resolutions that yield the least loss.

Regularization. To avoid overfitting (especially with small sample size) and achieve desirable properties in the learned parameters, we impose the following constraints to our model. We first assume that only a few edges in the graphs are highly associated with the variable of interest. This is a natural assumption as changes due to a brain disorder do not manifest over the whole brain but sparsely appear in different ROIs. We therefore impose an ℓ_1 -norm constraint to the first layer of \mathbf{W}^h which includes the fully connected weights. We expect that this constraint will set many elements in the first layer of \mathbf{W}^h to zeros and identify the edges that are highly related to the prediction of labels. Second, we expect \mathbf{s} to be smooth with an ℓ_2 -norm constraint. This lets us obtain a smoothly transiting multi-resolution representation, i.e., avoid \mathbf{s} from diverging.

With these assumptions and (10), we minimize our final objective function

$$\tilde{J}(\mathbf{s}, \mathbf{W}^h) = J(\mathbf{s}, \mathbf{W}^h) + \frac{\theta_1}{N_G} \|\mathbf{W}_1^h\|_1 + \frac{\theta_2}{N_G} \|\mathbf{s}\|_2 \quad (11)$$

where \mathbf{W}_1^h represents the weights of the first layer of DNN module, θ_1 and θ_2 are the regularization parameters for ℓ_1 -norm and ℓ_2 -norm respectively.

The MENET is trained based on the partial derivatives of the objective function (11) with respect to the trainable parameters, i.e., $\frac{\partial \tilde{J}(\mathbf{s}, \mathbf{W}^h)}{\partial w_{t,spq}^h}$ and $\frac{\partial \tilde{J}(\mathbf{s}, \mathbf{W}^h)}{\partial s}$.

These parameters \mathbf{W}^h and \mathbf{s} are then updated using gradient descent with different learning rates r_W and r_s respectively.

3 Experiments

In this section, we demonstrate experiments on two independent real datasets: structural brain connectivity from Diffusion Tensor Images (DTI) in Alzheimer’s Disease Neuroimaging Initiative (ADNI) and functional connectivity from resting-state functional magnetic resonance images (rs-fMRI) in ADHD-200.

3.1 Datasets

ADNI. From the initiative, individual DTIs were processed by our in-house tractography pipeline to extract structural brain networks using Destrieux atlas [7] with 148 ROIs. Each brain network is given as an adjacency matrix whose elements denote *number of neuron fiber tracts* connecting two different ROIs. The dataset included $N = 506$ subjects and we merged control (CN) and Early Mild Cognitive Impairment (EMCI) groups as *Pre-clinical AD* group and combined Late Mild Cognitive Impairment (LMCI) and AD groups as *Prodromal AD* group to ensure sufficient sample size and compare their subtle differences.

ADHD-200. We adopted rs-fMRI data which were registered to Automated Anatomical Labeling (AAL) atlas with 116 ROIs [33]. We computed Pearson’s correlation coefficients (without threshold) between 116 different ROIs to construct functional brain connectivity for each participant. Taking the samples without artifact and removing groups with few subjects, we ended up with total of $N = 756$ samples labeled as 1) Typically Developing Children (TDC), 2) ADHD-Combined (ADHD-C), and 3) ADHD-Inattentive (ADHD-I). Our result may vary from [2] as the experimental settings are different (Table 1).

Table 1. Demographics of ADNI and ADHD-200 Datasets

Category	ADNI		ADHD-200		
	Preclinical AD (CN,EMCI)	Prodromal AD (LMCI,AD)	TDC	ADHD-C	ADHD-I
# of Subjects	276 (109,167)	171 (94,77)	487	159	110
Age (mean, std)	72.7(73.8,72.0), 6.9(5.8,7.5)	74.2(72.6,76.1), 6.9(6.4,7.0)	12.2(3.3)	11.2(3.0)	12.0(2.6)
Gender (M/F)	163/113 (57/52, 106/61)	98/73 (51/43, 47/30)	258/229	130/29	85/25

3.2 Experimental Settings

Evaluation Measures. We used 3-fold cross validation (CV) to evaluate our model and baselines with unbiased results. Evaluation measures were accuracy, precision, recall and F1-score averaged across the folds.

Parameters. The kernel was defined as $k_s(x) = sxe^{-sx}$ to ensure that the $k()$ behaves as a band-pass filter (i.e., it achieves 0 at the origin), and the total number of scales to derive multi-resolution representation was $|s| = 5$. Weights were randomly initialized with Xavier initialization and the scales were uniformly selected between $[0.01, 2.5]$. The input to DL methods was the flattened \mathcal{L}_s and the number of hidden units was set to 256. LeakyReLU activation function with negative slope 0.2 and batch normalization were applied to hidden units. ℓ_1 -norm was applied on the first layer of DNN with hyper-parameter $\theta_1=0.0001$ to achieve sparsity. ℓ_2 -norm was adopted on scales with hyper-parameter $\theta_2=0.001$.

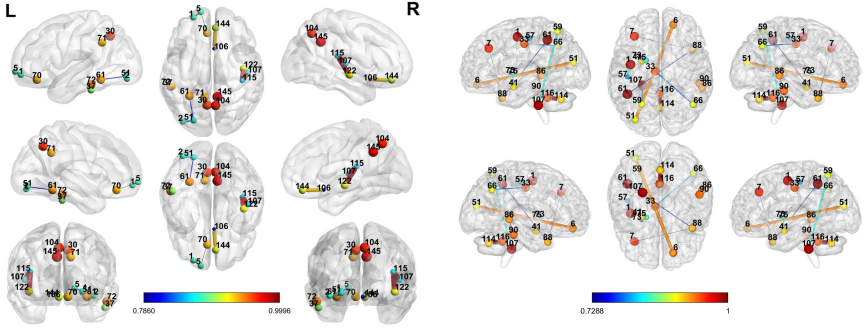


Fig. 3. Top-10 Connectivities from ADNI (Left) and ADHD-200 (Right) Analyses. Edge thickness denotes average trained edge weight and node color denotes its degree.

The learning rate of scale parameters was set to $r_s = 0.01$, and that of weight parameters was set to $r_W = 0.001$. The γ for focal loss was set to 2.

Baselines. We used SVM with RBF kernel, Logistic regression (LR), DNN, graph kernel (Shortest Path (SP)) with SVM, and several state-of-the-art graph learning frameworks (i.e., graph2vec [26] and DGCNN [35], and GCN [18]) as baselines for comparisons. For SVMs, we used Principal Component Analysis (PCA) with the rank of 10 to perform dimension reduction otherwise the models predicted all test samples as a single class with poor precision and recall. Node degree was used as node feature if a method required one.

3.3 Structural Brain Connectivity Analysis on ADNI

Result. Analysis on structural connectivity for Preclinical AD was performed. Binary classification task was designed to identify differences between the Preclinical ($N=276$) and Prodromal AD ($N=171$) groups instead of classifying four groups. As there is no effective treatment for AD, predicting the risk of developing dementia in the preclinical stage which is critical to enable disease-modifying interventions. Moreover, this task is particularly challenging since the physiological distinction between EMCI and LMCI, which lie along the decision boundary between Preclinical and Prodromal AD, is known to be especially subtle [4, 24].

The classification results of all baselines and MENET across four evaluation measures (accuracy, precision, recall, and F1-score) averaged across the folds are shown in Table 3. We show that MENET outperforms *all* other baselines. First, we see that SVM and graph kernel SVM achieved accuracy around $\sim 60\%$, close to 61.7% which is the random prediction would yield in this dataset, with poor precision and recall. GNNs (i.e., DGCNN and GCN) did not perform well either; this may be due to both GCN and DGCNN emphasizing nodes instead of edges that better characterize these particular graphs. While graph2vec with SVM showed higher performance than GNNs, its graph embedding is characterized

by node and not edges, hence not interpretable at the connectivity level. Interestingly, directly applying LR on these \mathcal{L} outperforms most of the previously mentioned baselines. This further demonstrates that the existing graph-based deep models may be *sub-optimal* for the current task; instead, the key is to effectively operate on \mathcal{L} as we observe next.

Ablation Study (\mathcal{L} vs. \mathcal{L}_s). DNN directly applied on \mathcal{L} immediately improves over LR. Although this gain is somewhat expected, we point out the significance of the multi-resolution graph transform \mathcal{L}_s : MENET, which is essentially DNN on \mathcal{L}_s with the *same* network classifier, shows significant improvements. MENET improves over DNN (*second best* model) in *all* evaluations by 10.9% in accuracy, 9.1% in precision, 14.6% in recall, and 14.0% in F1-score.

Clinical Validation. To help clinical interpretation of our findings, we investigate the trained edge weights that are connected to the trained \mathcal{L}_s , and visualize them as a brain network in Fig. 3. Only top-10 connectivities with the largest average weights are shown for sparsity—the most discriminative connectivities for Preclinical and Prodromal AD. The edge thickness and the node color respectively correspond to the edge weight and nodal degree of each ROI based on the trained weights. The list of the 10 connections is given along the figure that span across 17 ROIs (full ROI labels are in [7]). We observed several temporal regions (i.e., inferior temporal gyrus (37), inferior temporal sulcus (72), superior temporal gyrus (107)) [9, 16], precuneus (30, 104) [15], subparietal regions (71, 145) [5], and many others, all of which are corroborated by various AD literature.

3.4 Functional Brain Connectivity Analysis on ADHD

Result. Our results on multi-class classification on ADHD-200 are summarized in Table 3. We note that identifying the differences among these groups is innately challenging in ADHD-200 dataset consisting of adolescents with actively developing brains, inducing high variation. Further, compared to the previous AD experiment, this *multi-class* classification is particularly difficult especially with the severe class imbalance (64.4% random test set prediction accuracy). Thus, it is crucial that a method benefits all four evaluations in this analysis. In fact, throughout our experiments, we often observed undesirable behaviors

Table 2. List of Top-10 Significant Brain Connectivities from Fig. 3.

ADNI			ADHD-200		
Index	Row (ROI label [7])	Col (ROI label)	Index	Row(ROI label [33])	Col(ROI label)
1	115 (rLat_Fis-post)	107 (rG_temp_sup-G.T.transv)	1	114 (Vermis_8)	116 (Vermis_10)
2	122 (rS_circular_insula_inf)	115 (rLat_Fis-post)	2	6 (Frontal_Sup_Orb_R)	51 (Occipital_Mid_L)
3	106 (rG_subcallosal)	144 (rS_suborbital)	3	41 (Amygdala_L)	73 (Putamen_L)
4	51 (IS_collat_transv_post)	2 (IG_and_S_occipital_inf)	4	59 (Parietal_Sup_L)	107 (Cerebellum_10_L)
5	72 (IS_temporal_inf)	37 (IG_temporal_inf)	5	1 (Precentral_L)	57 (Postcentral_L)
6	51 (IS_collat_transv_post)	61 (IS_oc-temp_med_and_Lingual)	6	41 (Amygdala_L)	75 (Pallidum_L)
7	1 (IG_and_S_frontomargin)	5 (IG_and_S_transv_frontopol)	7	33 (Cingulum_Mid_L)	66 (Angular_R)
8	30 (IG_precuneus)	71 (IS_subparietal)	8	61 (Parietal_Inf_L)	88 (Tem_Pole_Mid_R)
9	145 (rS_subparietal)	104 (rG_precuneus)	9	86 (Temporal_Mid_R)	90 (Temporal_Inf_R)
10	5 (IG_and_S_transv_frontopol)	70 (IS_suborbital)	10	7 (Frontal_Mid_L)	88 (Tem_Pole_Mid_R)

Table 3. Classification Performances on ADNI and ADHD-200 Datasets.

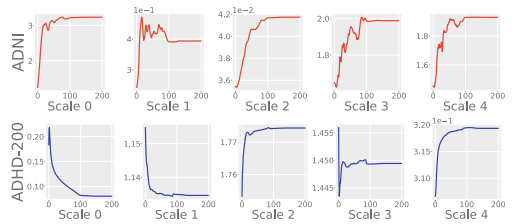
Model	ADNI				ADHD-200			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
SVM	62.19 \pm 5.86%	57.67 \pm 8.90%	56.93 \pm 7.17%	56.21 \pm 8.49%	52.06 \pm 4.39%	31.95 \pm 3.11%	32.56 \pm 2.28%	31.75 \pm 1.98%
LR	74.27 \pm 2.70%	74.47 \pm 1.91%	69.60 \pm 4.15%	70.08 \pm 4.63%	45.29 \pm 8.36%	27.18 \pm 4.56%	28.74 \pm 3.50%	27.21 \pm 3.53%
SP-SVM	58.61 \pm 1.14%	42.02 \pm 4.58%	48.13 \pm 1.14%	39.77 \pm 1.50%	51.13 \pm 4.25%	31.84 \pm 1.06%	32.36 \pm 1.00%	31.63 \pm 0.68%
DNN	78.65 \pm 0.92%	80.24 \pm 1.15%	74.49 \pm 2.35%	75.45 \pm 2.13%	56.04 \pm 7.17%	41.94 \pm 9.23%	34.63 \pm 0.96%	31.91 \pm 3.45%
Graph2vec	74.83 \pm 5.93%	75.06 \pm 5.56%	71.89 \pm 7.25%	71.69 \pm 7.64%	61.62 \pm 1.79%	21.36 \pm 0.08%	31.82 \pm 0.93%	25.56 \pm 0.35%
DGCNN	65.77 \pm 2.90%	73.20 \pm 11.48%	62.20 \pm 3.31%	61.89 \pm 4.03%	62.27 \pm 3.12%	37.47 \pm 13.34%	35.20 \pm 1.95%	29.94 \pm 4.06%
GCN	67.79 \pm 3.71%	67.67 \pm 7.30%	62.34 \pm 3.33%	62.21 \pm 3.44%	59.76 \pm 6.23%	39.68 \pm 10.91%	33.66 \pm 0.37%	29.49 \pm 3.08%
MENET	87.27 \pm 3.71%	87.56 \pm 2.90%	85.39 \pm 5.06%	86.02 \pm 4.41%	62.82 \pm 1.60%	44.95 \pm 4.46%	35.48 \pm 1.82%	32.41 \pm 3.38%

from these algorithms biasing towards predicting all testing instances as TDC for maximal high accuracy while sacrificing precision and recall. For instance, graph2vec+SVM achieved the third-highest accuracy with nearly the worst precision, recall and F1-score even with our best effort to alleviate this issue. Therefore, all the models were carefully tuned to prevent such cases as much as possible. Again, MENET achieved higher performance than all baseline models in *all* evaluation metrics by achieving the highest average accuracy at 62.82% while also improving precision, recall and F1-score. DNN, DGCNN and GCN suffered from large standard deviations, indicating unstable learning across the CV folds (Table 2).

Clinical Validation. To help clinical interpretation of our findings, we similarly derived the top-10 functional connectivities that are associated with the classification of ADHD stages. The top-10 connectivities with the highest average weights are shown in Fig. 3, whose edge thickness and node color correspond to the edge weight and degree of each ROI respectively. The list of the 10 connections, that span across 18 ROIs, is given along the figure (full ROI names in [33]). We observed several ROIs in temporal regions (mid-temporal pole (86, 88), inferior temporal cortex (90)) [29], frontal regions (frontal superior orbital (6), mid-frontal (7)) [25], and Amygdala (41) [23]; these results along with other identified ROIs are already well documented in many ADHD literature.

3.5 Discussions on Convergence of Scales

Figure 4 shows the convergence of scales for AD (top) and ADHD (bottom) experiments. The scale parameters of our model converge very fast; the low-pass filter (i.e., Scale 0) of the ADNI experiment tends to increase while ADHD tends to decrease in the experiments. One possible reason is that the ADHD classification requires the model to involve more local context information as the problem is more difficult than the AD experiment.

**Fig. 4.** Convergence of scales w.r.t. training epoch. Top: ADNI, Bottom: ADHD-200.

4 Conclusion

We developed a novel graph transform-based CNN framework designed to perform classification tasks with a population of registered graphs. The transform derives multi-resolution representations of a graph matrix, i.e., edges, that serve as effective features suited to perform classification on graphs. Using a parametric kernel, our framework, i.e., MENET, can train well with relatively small sample size and was validated with extensive experiments on two independent connectivity datasets, yielding clinically sound results on AD and ADHD supported by existing literature. We believe that MENET has significant potential to domains with graph data practically challenged by small sample sizes.

Acknowledgments. This research was supported by NSF IIS CRII 1948510, NSF IIS 2008602, NIH R01 AG059312, IITP-2020-2015-0-00742, and IITP-2019-0-01906 funded by MSIT (AI Graduate School Program at POSTECH).

References

1. Borgwardt, K.M., Kriegel, H.P.: Shortest-path kernels on graphs. In: ICDM (2005)
2. Brown, M.R., Sidhu, G.S., Greiner, R., et al.: ADHD-200 global competition: diagnosing ADHD using personal characteristic data can outperform resting state fMRI measurements. *Front. Syst. Neurosci.* **6**, 69 (2012). <https://doi.org/10.3389/fnsys.2012.0006>
3. Bullmore, E., Sporns, O.: Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* **10**(3), 186 (2009)
4. Chincarini, A., Bosco, P., Calvini, P., et al.: Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. *Neuroimage* **58**(2), 469–480 (2011). <https://doi.org/10.1016/j.neuroimage.2011.05.083>
5. Choo, I.H., Lee, D.Y., Oh, J.S., et al.: Posterior cingulate cortex atrophy and regional cingulum disruption in mild cognitive impairment and Alzheimer's disease. *Neurobiol. Aging* **31**(5), 772–779 (2010). <https://doi.org/10.1016/j.neurobiolaging.2008.06.015>
6. Dennis, E.L., Thompson, P.M.: Functional brain connectivity using fMRI in aging and Alzheimer's disease. *Neuropsychol. Rev.* **24**(1), 49–62 (2014). <https://doi.org/10.1007/s11065-014-9249-6>
7. Destrieux, C., Fischl, B., Dale, A., Halgren, E.: Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53**(1), 1–15 (2010). <https://doi.org/10.1016/j.neuroimage.2010.06.010>
8. Friston, K.J.: Statistical Parametric Mapping. In: Kötter, R. (eds.) *Neuroscience Databases*. Springer, Boston (2003). https://doi.org/10.1007/978-1-4615-1079-6_16
9. Galton, C.J., Patterson, K., Graham, K., et al.: Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurology* **57**(2), 216–225 (2001). <https://doi.org/10.1212/wnl.57.2.216>
10. Gärtner, T., Flach, P., Wrobel, S.: On graph kernels: hardness results and efficient alternatives. In: Schölkopf, B., Warmuth, M.K. (eds.) *COLT-Kernel 2003*. LNCS (LNAI), vol. 2777, pp. 129–143. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-45167-9_11

11. Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V.: Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci.* **100**(1), 253–258 (2003). <https://doi.org/10.1073/pnas.0135058100>
12. Hammond, D., Vandergheynst, P., Gribonval, R.: Wavelets on graphs via spectral graph theory. *Appl. Comput. Harmon. Anal.* **30**(2), 129–150 (2011). <https://doi.org/10.1016/j.acha.2010.04.005>
13. Jie, B., Zhang, D., Wee, C.Y., Shen, D.: Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification. *Hum. Brain Mapp.* **35**(7), 2876–2897 (2014). <https://doi.org/10.1002/hbm.22353>
14. Kang, U., Tong, H., Sun, J.: Fast random walk graph kernel. In: *ICDM* (2012)
15. Karas, G., Scheltens, P., Rombouts, S., et al.: Precuneus atrophy in early-onset Alzheimer's disease: a morphometric structural MRI study. *Neuroradiology* **49**(12), 967–976 (2007). <https://doi.org/10.1007/s00234-007-0269-2>
16. Kim, W.H., Pachauri, D., Hatt, C., et al.: Wavelet based multi-scale shape features on arbitrary surfaces for cortical thickness discrimination. *Adv. Neural Inf. Process. Syst.* **2012**, 1241–1249 (2012)
17. Kim, W.H., Kim, H.J., Adluru, N., Singh, V.: Latent variable graphical model selection using harmonic analysis: applications to the human connectome project (hcp). In: *CVPR*, pp. 2443–2451 (2016)
18. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: *ICLR* (2017)
19. Landrieu, L., Simonovsky, M.: Large-scale point cloud semantic segmentation with superpoint graphs. In: *CVPR*, pp. 4558–4567 (2018)
20. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: *ICCV*, pp. 2980–2988 (2017)
21. Ma, X., Wu, G., Kim, W.H.: Enriching statistical inferences on brain connectivity for Alzheimer's disease analysis via latent space graph embedding. In: *ISBI*, pp. 1685–1689. *IEEE* (2020)
22. Mallat, S.: *A Wavelet Tour of Signal Processing*. Academic Press (1999)
23. Marsh, A.A., Finger, E.C., Mitchell, D.G., et al.: Reduced amygdala response to fearful expressions in children and adolescents with callous-unemotional traits and disruptive behavior disorders. *Am. J. Psychiatry* **165**(6), 712–720 (2008). <https://doi.org/10.1176/appi.ajp.2007.07071145>
24. Moradi, E., Pepe, A., Gaser, C., Huttunen, H., Tohka, J., Initiative, A.D.N.: Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage* **104**, 398–412 (2015). <https://doi.org/10.1016/j.neuroimage.2014.10.002>
25. Murias, M., Swanson, J.M., Srinivasan, R.: Functional connectivity of frontal cortex in healthy and ADHD children reflected in EEG coherence. *Cereb. Cortex* **17**(8), 1788–1799 (2006)
26. Narayanan, A., Chandramohan, M., Venkatesan, R., et al.: graph2vec: learning distributed representations of graphs. *arXiv preprint arXiv:1707.05005* (2017)
27. Ng, B., Varoquaux, G., Poline, J.B., Thirion, B., Greicius, M.D., Poston, K.L.: Distinct alterations in Parkinson's medication-state and disease-state connectivity. *NeuroImage Clin.* **16**, 575–585 (2017)
28. Riesen, K., Bunke, H.: Graph classification based on vector space embedding. *Int. J. Pattern Recognit. Artif. Intell.* **23**(06), 1053–1081 (2009). <https://doi.org/10.1142/7731>

29. Rubia, K., Smith, A.B., Brammer, M.J., Taylor, E.: Temporal lobe dysfunction in medication-Naive boys with attention-deficit/hyperactivity disorder during attention allocation and its relation to response variability. *Biol. Psychiat.* **62**(9), 999–1006 (2007). <https://doi.org/10.1016/j.biopsych.2007.02.024>
30. Shervashidze, N., Schweitzer, P., Van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-Lehman graph kernels. *JMLR* **12**(9), 2539–2561 (2011)
31. Shervashidze, N., Vishwanathan, S., Petri, T., Mehlhorn, K., Borgwardt, K.: Efficient graphlet kernels for large graph comparison. In: *AISTATS* (2009)
32. Simonovsky, M., Komodakis, N.: Dynamic edge-conditioned filters in convolutional neural networks on graphs. In: *CVPR*, pp. 3693–3702 (2017)
33. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., et al.: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**(1), 273–289 (2002). <https://doi.org/10.1006/nimg.2001.0978>
34. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017)
35. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph CNN for learning on point clouds. *ACM Trans. Graph.* **38**, 146 (2019)
36. Xu, N., Wang, P., Chen, L., Tao, J., Zhao, J.: MR-GNN: multi-resolution and dual graph neural network for predicting structured entity interactions. *arXiv preprint arXiv:1905.09558* (2019)
37. Zhang, M., Cui, Z., Neumann, M., Chen, Y.: An end-to-end deep learning architecture for graph classification. In: *AAAI* (2018)