Medical Image Analysis 48 (2018) 117–130

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Disease prediction using graph convolutional networks: Application to Autism Spectrum Disorder and Alzheimer's disease

Sarah Parisot[b,1,2,*], Sofia Ira Ktena[a,1], Enzo Ferrante[c], Matthew Lee[a], Ricardo Guerrero[d,3], Ben Glocker[a], Daniel Rueckert[a]

[a] Biomedical Image Analysis Group, Imperial College London, UK
[b] AimBrain Solutions Ltd, London, UK
[c] Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL/CONICET, Santa Fe, Argentina
[d] StoryStream Ltd., London, UK

ARTICLE INFO

Article history:
Received 2 February 2018
Revised 31 May 2018
Accepted 1 June 2018
Available online 2 June 2018

Keywords:
Graphs
Graph convolutional networks
Spectral theory
Semi-supervised classification
Autism Spectrum Disorder
Alzheimer's disease

ABSTRACT

Graphs are widely used as a natural framework that captures interactions between individual elements represented as nodes in a graph. In medical applications, specifically, nodes can represent individuals within a potentially large population (patients or healthy controls) accompanied by a set of features, while the graph edges incorporate associations between subjects in an intuitive manner. This representation allows to incorporate the wealth of imaging and non-imaging information as well as individual subject features simultaneously in disease classification tasks. Previous graph-based approaches for supervised or unsupervised learning in the context of disease prediction solely focus on pairwise similarities between subjects, disregarding individual characteristics and features, or rather rely on subject-specific imaging feature vectors and fail to model interactions between them. In this paper, we present a thorough evaluation of a generic framework that leverages both imaging and non-imaging information and can be used for brain analysis in large populations. This framework exploits Graph Convolutional Networks (GCNs) and involves representing populations as a sparse graph, where its nodes are associated with imaging-based feature vectors, while phenotypic information is integrated as edge weights. The extensive evaluation explores the effect of each individual component of this framework on disease prediction performance and further compares it to different baselines. The framework performance is tested on two large datasets with diverse underlying data, ABIDE and ADNI, for the prediction of Autism Spectrum Disorder and conversion to Alzheimer's disease, respectively. Our analysis shows that our novel framework can improve over state-of-the-art results on both databases, with 70.4% classification accuracy for ABIDE and 80.0% for ADNI.

1. Introduction

Large scale collaborative initiatives and consortiums, like the Alzheimer's Disease Neuroimaging Initiative (ADNI), the Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium (Thompson et al., 2014) and the International Neuroimaging Data-sharing Initiative (INDI),[4] acquire and share hundreds of terabytes of imaging, genetic, phenotypic and behavioural data in an effort to facilitate the discovery of novel biomarkers and better understand disease mechanisms. This ever-increasing volume of imaging and non-imaging information that is collected and shared among researchers stresses the need for computational models that are capable of representing potentially large populations, while exploiting all imaging modalities and additional data sources available.

Graphs provide a powerful and intuitive way of modelling individuals (as nodes) and associations or similarities between them (as edges). In this setting, a node can represent a specific modality of a subject's acquired data or a series of acquisitions at a particu-

* Corresponding author.
E-mail address: sarah@aimbrain.com (S. Parisot).
[1] Authors contributed equally.
[2] Aimbrain Solutions ltd, One Canada Square, London, E14 5AB. This work was carried out when the first author was affiliated with the Biomedical Image Analysis Group, Imperial College London, UK.
[3] This work was carried out when the author was affiliated with the Biomedical Image Analysis Group, Imperial College London, UK.

[4] http://fcon_1000.projects.nitrc.org/

https://doi.org/10.1016/j.media.2018.06.001
1361-8415/

lar time point, while the edge weights can be used to capture the similarities between each pair of nodes. Edges can also be non-weighted in this context, or binarised to indicate the absence or presence of association between two subjects/scans, which, however, limits their expressiveness. Graph-based models have been widely used for supervised (e.g. classification (Tong et al., 2017b)) and unsupervised tasks (e.g. manifold learning (Wolz et al., 2012; Brosch and Tam, 2013) and clustering (Parisot et al., 2016)) in population analysis, due to their capacity to accommodate complex pairwise interactions and ability to integrate non-imaging information. However, previous approaches have focused on summarising all available feature information via pairwise similarities and thus eliminated individual subject features and characteristics. For example, Tong et al. (2017b) proposed a non-linear population graph fusion approach that aims to combine complementary multi-modal information about the subjects, while Zhao et al. (2014) focused on the graph construction and proposed an improved local reconstruction strategy for graph-based label propagation in Alzheimer's disease classification. On the contrary, methods that solely rely on imaging feature vectors (Cuingnet et al., 2011; Abraham et al., 2017) by training, for example, linear classifiers, are widely applied for classification analyses in large populations, but fail to capture interactions and similarities between subjects or their individual scans. A progressive method that seeks to learn an affinity matrix from the observed imaging features while validating it on the training data with known phenotype labels is proposed in Wang et al. (2017), aiming to augment imaging with phenotypic information. Such analyses become more challenging and limited in performance when diverse imaging protocols are adopted for data acquisition within the same population, since they are more difficult to generalise.

1.1. Graph convolutional neural networks

The advent of Convolutional Neural Networks (CNNs) as powerful models that exploit both image features (e.g. intensities) and spatial context by means of neighbourhood information (e.g. regular pixel grid) to yield hierarchies of features, has led to their application in numerous different problems related to 2D and 3D images, like image segmentation (Havaei et al., 2017) and classification (Hou et al., 2016), long before their recent re-emergence (Sahiner et al., 1996). There is a direct analogy between an image segmentation task, where each pixel is to be assigned a label e.g. tissue of interest or background, and a subject classification task within a population e.g. for disease prediction. In the latter case, a subject along with its corresponding feature vector is equivalent to an image pixel with its intensity values for the different channels, while a graph constructed based on pairwise population similarities is equivalent to the pixel grid, for which the notion of proximity is more straightforward, since both describe the neighbourhood structure for convolutions. Nevertheless, the traditional widespread formulation of CNNs for regular domains cannot be directly extended to irregular ones. The description of the local neighbourhood structure and node ordering is not straightforward for irregular graph structures and these need to be properly defined to allow convolution and pooling operations (Niepert et al., 2016).

The first method dealing with neural networks on graphs was presented in Scarselli et al. (2009). In this pioneering work, the authors devised a mapping function from graph space to an m-dimensional Euclidean space, and proposed a supervised learning method to learn the parameters of their graph neural network (GNN) model. However, no convolution was considered in this model. The first work to introduce convolutional neural networks on graphs was described in Bruna et al. (2013). Bruna et al. used concepts from the emerging field of signal processing on graphs (Shuman et al., 2013), giving rise to the spectral graph convolutional networks (GCNs). Spectral GCN is a generalisation of CNNs to irregular domains, which uses computational harmonic analysis to process signals observed on irregular graph structures (Hammond et al., 2011). The concepts borrowed from this field allow the extension of CNNs to irregular graphs in a principled way, by treating convolutions in the graph spatial domain as multiplications in the graph spectral domain. This kind of convolutions have several applications in computer vision, computer graphics and social network problems, among others, and have successfully been adopted to perform classification of documents in large citation datasets (Kipf and Welling, 2016; Levie et al., 2017). Alternative approaches, where convolutions are directly defined in the spatial domain, have also been proposed in the literature. In Masci et al. (2015), for example, the authors define a local geodesic system of polar coordinates to extract patches, which are processed through a cascade of filters and activation functions. The so called geodesic convolutions are then used to construct graph convolutional networks which operate directly on the manifold. Another recent work by Simonovsky and Komodakis (2017) proposed to use filter weights conditioned on the edge labels, instead, and dynamically generate those for each input sample. This type of spatial convolutions on graphs are especially suitable for mesh structures, since they are local and allow to capture anisotropic patterns. A more detailed overview of such techniques is presented in Bronstein et al. (2017). However, spectral convolutions are preferential in cases where the underlying graph structure is fixed, as spatial graph CNNs tend to require more engineering. In this work, we use the convolution approach described in Defferrard et al. (2016), since it has shown outstanding performance for node classification tasks where the graph structure models different types of interactions between individuals within a population (Kipf and Welling, 2016).

1.2. Graph-based models for disease prediction

Recently, the use of graph-based models has gained popularity in medical imaging applications, especially at a subject level. Kawahara et al. (2017) used a customised version of CNNs on structural connectivity matrices to predict neurodevelopmental outcomes in preterm infants, which operates on the graph spatial domain and, therefore, captures only 1-hop neighbours in the receptive field of each node. As an alternative, spectral methods have been used to learn a similarity metric between functional connectivity networks (Ktena et al., 2017), which was applied for Autism Spectrum Disorder (ASD) and sex classification as well as manifold learning (Ktena et al., 2018). Spectral methods have also been explored for the prediction of visual tasks from MEG signals on a small number of subjects (Guo et al., 2017), while a bootstrapping strategy was used by Anirudh and Thiagarajan (2017) for ASD prediction. Finally, Lombaert et al. (2015) combine spectral theory with random forests to process brain surfaces using spectral representations of meshes. Their proposed Spectral Forests are applied to the brain parcellation problem.

In Parisot et al. (2017), we proposed the first application of GCNs for group-level medical applications, more specifically for brain analysis in populations and diagnosis. We modelled populations as sparse graphs, where each node represents a subject and is associated with a feature vector extracted from imaging data. The edge weights encode the pairwise similarities between subjects and their features, and are obtained from auxiliary phenotypic data. This enables us to combine imaging and non-imaging data in a single framework. This population graph is fed as input to a GCN, which is trained in a semi-supervised manner from a subset of labelled nodes (e.g. of known diagnosis), aiming to classify the remaining, unlabelled nodes. The goal of the proposed method

is to leverage the complementary non-imaging information available to explain the similarities between subjects within a graph structure and exploit the power of graph convolutions. Our main hypothesis is that integrating clinical expertise, i.e. non-imaging information that is known to be linked to specific pathologies, to model similarities between subjects can improve learned representations of image features and classification performance.

Contribution: This paper constitutes an extended version of our work in Parisot et al. (2017). In this work, we provide a deeper analysis of the method and modelling choices, through a substantially extended experimental evaluation as well as an in-depth discussion. We explore the influence of each main component of the model and provide discussion on the implications of the obtained results. This extensive evaluation is carried out on the ABIDE and ADNI databases, two large and challenging datasets, with the aim to diagnose Autism Spectrum Disorder (ABIDE) and to predict conversion from Mild Cognitive Impairment (MCI) to Alzheimer's Disease (ADNI). Our evaluation on two different datasets aims to demonstrate the framework's versatility as it facilitates the incorporation of domain-specific knowledge in two different clinical settings, while at the same time showing consistent improvement with respect to baselines for both challenging problems.

The main contributions of the proposed work are:

- An introduction of GCNs for population analysis in the medical imaging domain.
- A novel formulation of subject classification as a graph labelling problem, integrating imaging and non imaging data.
- A seamless integration of known non-imaging features, allowing to integrate clinical expertise to boost classification performance.

In addition, our extended version proposes the following enhancements with respect to the work introduced in Parisot et al. (2017):

- A complete sensitivity analysis for key parameters of the proposed GCN model
- New feature selection strategies for the ABIDE database
- Detailed investigation of different graph structures and their influence on classification results
- Comparison of our method to new baselines: Random Forest and Multi-Layer Perceptron classifiers
- An improved, state of the art performance on both databases

Our experiments show that exploiting GCNs with an accurate graph structure leads to significant improvements in classification accuracy, yielding state of the art results Arbabshirani et al. (2017), Abraham et al. (2017) and Heinsfeld et al. (2018) for both ABIDE (70.4% accuracy) and ADNI (80% accuracy). The model implementation is publicly available at https://github.com/parisots/population-gcn.

2. Methods

An overview of the proposed method is shown in Fig. 1. We consider a population of S subjects, each subject being described by/associated with a set of complimentary, phenotypic and demographic information (e.g. sex, age, acquisition site). The population comprises a set of N imaging acquisitions (structural or functional MRI are considered in this paper), with $N \geq S$, meaning that one subject can be associated with several acquisitions (longitudinal scans). Our objective is to predict the status of each subject (healthy control or disease) from the imaging data informed by the phenotypic information. The population is represented as a weighted sparse graph, $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, W\}$, where W is the adjacency matrix describing the graph's connectivity. Each acquisition A_v is represented by a vertex $v \in \mathcal{V}$, corresponds to a subject S_v and is associated with a C-dimensional feature vector $\mathbf{x}(v)$ extracted from the imaging data. The edges $\mathcal{E}$ of the graph model the similarity between the corresponding subjects and incorporate the phenotypic information.

We model our diagnosis task as a node classification problem, where we aim to assign a label $l \in \{0, 1\}$ to each graph node which describes the diseased ($l = 1$) or healthy status ($l = 0$) of the subject. Even though we focus on binary classification in this work, the model can easily be extended for multi-class classification problems. We adopt a semi-supervised strategy, where all node features along with the population graph are fed to the GCN, while only a subset of the graph nodes is labelled during training and used for the optimisation process. Intuitively, the graph acts as a regulariser, 'encouraging' nodes connected with high edge weights to contribute more towards filtering the features of neighbouring nodes in a way that boosts label propagation performance.

2.1. Databases and preprocessing

We demonstrate the potential and versatility of the model using two large and challenging databases, namely the *ABIDE* and *ADNI* databases. Each of the databases comprises subject specific information that can be leveraged through our population graph structure.

The **ABIDE database** (Di Martino et al., 2014) aggregates data from different international acquisition sites and openly shares neuroimaging (functional MRI) and phenotypic data of 1112 subjects.[5] We select the same set of 871 subjects used by Abraham et al. (2017) that met the imaging quality and phenotypic information criteria, comprising 403 individuals with ASD and 468 healthy controls. These subjects were regrouped based on location into 20 imaging sites. To ensure a fair comparison with the state of the art (Abraham et al., 2017) we use the same preprocessing pipeline, the Configurable Pipeline for the Analysis of Connectomes (C-PAC) (Craddock et al., 2013), which involves skull striping, slice timing correction, motion correction, global mean intensity normalisation, nuisance signal regression, band-pass filtering (0.01-0.1Hz). The functional images were registered to a standard anatomical space (MNI152) to allow cross-subject comparisons. Subsequently, the mean time series for a set of regions extracted from the Harvard Oxford (HO) atlas (Desikan et al., 2006) were computed and normalised to zero mean and unit variance. The HO atlas distributed with FSL[6] is based on anatomical landmarks and the version used in this work includes both cortical and subcortical (excluding left/right WM, left/right GM, left/right CSF and brainstem) ROIs, yielding 111 regions in total. The individual connectivity matrices $M_1, \ldots, M_N$ are estimated by computing the Fisher transformed Pearson's correlation coefficient between the representative rs-fMRI timeseries of each ROI in the HO atlas. The correlation matrices are, then, Fisher transformed to improve normality.

The **ADNI database** is the result of efforts from several academic and private co-investigators.[7] To date, ADNI in its three studies (ADNI-1, -GO and -2) has recruited over 1700 adults, aged between 55 and 90 years, from over 50 sites from the U.S. and Canada. In this work, a subset of 540 early/late MCI subjects that contained longitudinal T1 MR images and their respective anatomical segmentations was used. Our inclusion criteria were therefore: MCI diagnosis, available longitudinal T1 MR images and available corresponding segmentations as described in Ledig et al. (2015). MCI often represents an intermediate stage between normal cognition and Alzheimer's disease. Therefore, the conversion from MCI

[5] http://preprocessed-connectomes-project.org/abide/ .
[6] http://www.fmrib.ox.ac.uk/fsl/.
[7] http://adni.loni.usc.edu.

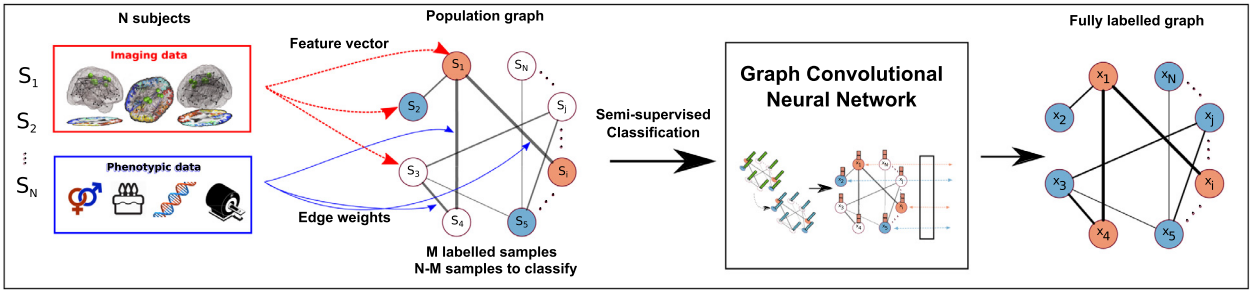


Fig. 1. Overview of the pipeline used for classification of population graphs using Graph Convolutional Networks.

to AD is more challenging to predict than distinguishing between HC and AD patients. In total, 1675 samples were available, with 289 subjects (843 samples) diagnosed as AD at any time during follow-up and labelled as converters. It should be noted that the AD diagnosis in the dataset comes from clinical evaluation only and is missing histopathological confirmation. As a result, the AD diagnosis should be referred to as "probable AD". For simplification purposes, we use the term AD to refer to probable AD in the remainder of this paper. Longitudinal information ranged from 6 to 96 months, depending on each subject. Acquisitions after conversion to AD were not included. As of 1st of July 2016 the ADNI repository contained 7128 longitudinal T1 MR images from 1723 subjects. ADNI-2 is an ongoing study and therefore data is still growing. Therefore, at the time of a large scale segmentation analysis (into 138 anatomical structures using MALP-EM (Ledig et al., 2015)) only a subset of 1674 subjects (5074 images) was processed, from which the subset used here was selected.

With the *ABIDE database*, we aim to separate healthy controls from patients suffering from ASD. This database comprises data acquired at different sites and using different protocols, which results in a highly heterogeneous database where the acquisition protocol can strongly affect the comparability of subjects. Our goal with the *ADNI database* is to predict whether a patient with MCI will convert to AD, in other words, we aim to separate patients with stable MCI from those with progressive MCI. Our strategy is to intrinsically model/exploit the longitudinal aspect of the data within our graph structure, so as to highlight the importance and advantage of modelling the interactions between different input data.

2.2. Population graph construction

We provide an illustration of the construction of the population graph in Fig. 1. This graph construction is a key aspect of the method, as an inappropriately constructed graph (i.e. a graph that does not accurately explain the similarity between subjects and their feature vectors) will fail to exploit the power of GCNs. Very inaccurate graph structures could even worsen performance compared to a simple linear classifier. Intuitively, this would equate to performing image convolutions on unrelated pixels (e.g. that are randomly spread across the image) instead of a local image patch. The two main decisions required to build the population model are : 1) the definition of the feature vector $\mathbf{x}(v)$ describing each graph node/acquisition, and 2) the connectivity of the graph, i.e. its edges $\mathcal{E}$ and their weights W, which models the similarity between nodes/subjects/scans and their corresponding features.

2.2.1. Feature vector

We extract the feature vector purely from imaging data, as would typically be the case for image based classification. Our objective is to demonstrate how classification using simple image features can be enhanced using complementary information in the graph structure.

For the ADNI dataset, we use the volumes of all $C = 138$ segmented brain structures, a type of feature which has been highly effective for prediction of Alzheimer's disease, due to the impact of the disease on the brain's structure and volume differences between healthy, MCI and AD populations (Ries et al., 2008).

There is increasing evidence that ASD is linked to disruptions in the functional and structural organisation of the brain Abraham et al. (2017) and Rudie et al. (2013). As a result, we use functional connectivity derived from resting-state functional Magnetic Resonance Imaging (rs-fMRI) for ASD classification using the ABIDE dataset. More specifically, we use the vectorised functional connectivity matrices, i.e. the upper triangular elements of the square adjacency matrices, as feature vectors. This simple approach has had numerous successes for fMRI based classification, it was notably used in Abraham et al. (2017), setting the state of the art on the whole ABIDE dataset at 67% with a simple linear classifier. Due to the high dimensionality of the connectivity matrix, we explore, alongside using the whole feature vector, different dimensionality reduction strategies to input a C dimensional feature vector to the network. Our different strategies, detailed in Section 2.3, are recursive feature elimination using a ridge classifier, a simple autoencoder, a multilayer perceptron classifier and principal component analysis. It is worth noting that feature selection is not used for ADNI data, which has a much smaller and tractable feature vector size.

2.2.2. Graph edges

Similarly to pixel neighbourhood systems, the graph structure provides a broader field of view, filtering the value of a feature with respect to its neighbours' instead of treating each feature individually. It has to be carefully crafted so as to accurately model the interactions between feature vectors. Our hypothesis is that non-imaging complementary data can provide key information to explain the associations between subjects' feature vectors. The objective is to leverage this information, in order to define an accurate neighbourhood system that optimises the performance of the subsequent graph convolutions. Therefore, it is important to select the phenotypic measures which best explain similarities between the imaging data, or similarities between the subjects' labels.

Considering a set of H non-imaging phenotypic measures $\mathbf{M} = \{M_h\}$ (e.g. subject's sex, or age), the population graph's adjacency matrix W is defined as follows:

$$W(v, w) = Sim(A_v, A_w) \sum_{h=1}^{H} \gamma(M_h(v), M_h(w)), \tag{1}$$

where, $Sim(S_v, S_w)$ is a measure of similarity between subjects, increasing the edge weights between the most similar graph nodes; γ is a measure of distance between phenotypic measures.

γ is defined differently depending on the type of phenotypic measure integrated in the graph. For categorical information such as subject's sex, we define γ as the Kronecker delta function δ, meaning that the edge weight between subjects is increased if e.g.

they have the same sex. Constructing edge weights from quantitative measures (e.g. subject's age) is slightly less straightforward. In such cases, we define γ as a unit-step function with respect to a threshold θ:

$$\gamma(M_h(v), M_h(w)) = \begin{cases} 1 & \text{if } |M_h(v) - M_h(w)| < \theta \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Both ADNI and ABIDE databases provide an extensive list of phenotypic features. In this work, we select three different phenotypic measures for each database that are considered relevant to the corresponding disease. For ABIDE, we consider *acquisition site, sex* and *age* as our three potential phenotypic measures. The ABIDE dataset is highly heterogeneous due to the fact that the data has been acquired at different sites, using diverse imaging protocols and scanners at each location. As a result, the imaging data is best comparable between feature vectors acquired at the same site, making acquisition site an essential non-imaging measure to introduce. Sex and age are also considered, since sex differences have been observed in several studies on ASD suggesting that females are affected less frequently than males (Werling and Geschwind, 2013), while there are age-related group differences in functional connectivity overall (Kana et al., 2014). Finally, we define the similarity measure as

$$Sim(A_v, A_w) = \exp\left(-\frac{[\rho(\mathbf{x}(v), \mathbf{x}(w))]^2}{2\sigma^2}\right), \quad (3)$$

where ρ is the correlation distance and σ determines the width of the kernel. The idea behind this similarity measure is that subjects belonging to the same class (healthy or ASD) tend to have more similar networks (larger *Sim* values) than subjects from different classes.

The ADNI graph is also built using the subject's *sex* and *age* information. These measures are chosen because our feature vector comprises brain volumes, which can strongly be affected by age and sex. We also integrate genetic information, namely the presence of the APOE $\epsilon 4$ allele, known to be a major risk factor for the development of Alzheimer's disease (Mosconi et al., 2004).

Last but not least, the $Sim(A_v, A_w)$ function plays a particularly important role in the ADNI set-up. It is designed here to leverage the longitudinal information, strongly highlighting that several acquisitions correspond to the same subject. While linear classifiers treat each entry independently, here we define:

$$Sim(A_v, A_w) = \lambda \text{ with } \begin{cases} \lambda > 1 & \text{if } S_v = S_w \\ \lambda = 0 & \text{otherwise.} \end{cases} \quad (4)$$

This measure indicates the strong similarity between acquisitions of the same subject.

The influence of each phenotypic and similarity measure will be investigated in our experiments section, so as to optimise the structure of our phenotypic graph. It should be noted that while handedness constitutes an important phenotypic measure associated to autism subtypes (Soper et al., 1986), we could not integrate it in this study as this information was missing for a large number of subjects.

2.3. Feature selection strategies

As mentioned in Section 2.2, the feature vector chosen for ABIDE classification (the vectorised fMRI connectivity network) has a high dimensionality (particularly with respect to the graph size of 871 nodes) which can negatively impact the performance of the algorithm and lead to overfitting issues. In Parisot et al. (2017), we used a ridge classifier to perform **Recursive Feature Elimination** (RFE) with a fixed number of features C. It is an iterative process, where each iteration trains the ridge classifier on the training set with the current feature vector. The classifier's coefficients are used to sort the importance of the features and prune the ones with the smallest coefficient (i.e. the least discriminative) from the feature vector. This process iterates until the desired number of features is obtained. In this paper, we investigate 3 additional approaches as well as the influence of C.

First, we use the well known **Principal Component Analysis** (PCA), which is very commonly used for dimensionality reduction purposes, using singular value decomposition to project the data to a space of lower dimensionality. PCA is not very adapted to our problem due to the much larger dimension of the feature vector compared to the number of samples (6105 vs 871).

Second, we use two simple models based on artificial neural networks, one supervised and one unsupervised. The first approach is a **Multilayer Perceptron** (MLP), a supervised feedforward neural network which consists of one hidden layer of size C. It is trained as a classifier using the ABIDE training data. The feature vector is obtained by extracting the C dimensional feature vector obtained at the MLP's hidden layer. The underlying idea is that the MLP will learn a representation of the data specifically for the classification task, similar to what is done using the RFE approach. However, even if we are using a shallow MLP with a single hidden layer, there is a strong possibility of overfitting due to our limited amount of training data. The MLP model is illustrated in Fig. 2.

Autoencoders are unsupervised neural networks that aim to learn a lower dimensional representation (a code) of an input data. It is made of an encoder (which learns the code) and a decoder (which reconstructs the input). It is trained by comparing the reconstructed input to the original data. Our autoencoder has tied weights and comprises one single hidden layer of size C with a sigmoid activation and a tanh activation at the output layer. We use the mean square error as a loss function for training. The autoencoder model is illustrated in Fig. 3.

2.4. Graph labelling using graph convolutional neural networks

In the previous section we described the construction of the population graph based on phenotypic measures, and feature selection strategies on the original imaging data. In this section, we present the concept of spectral graph convolutions that serves as building block for the final graph convolutional neural network model, and discuss the GCN's architectural details.

2.4.1. Spectral graph convolutions

The discretised convolutions commonly used in computer vision intrinsically exploit the regular grid-like structure of e.g. 2D or 3D images. As a result, their generalisation/application to irregular graphs is not straightforward. We propose to use a recent formulation relying on spectral theory and graph signal processing (Shuman et al., 2013).

Spatial graph convolutions can be computed in the Fourier (spectral) domain as multiplications. The concept of graph Fourier transform (GFT) is introduced in Shuman et al. (2013) by analogy with the Euclidean domain, as an expansion of the Laplace operator in terms of its eigenfunctions. The normalised graph Laplacian of a weighted graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, W\}$ is defined as $\mathcal{L} = I_N - D^{-1/2}WD^{-1/2}$ where I_N and D are respectively the identity matrix of size $N \times N$ and the diagonal degree matrix. For any signal $\mathbf{x} \in \mathbb{R}^N$ the graph Laplacian acts as a difference operator and yields

$$(\mathcal{L}x)(i) = \sum_{j \in \mathcal{N}_i} W_{ij}(x(i) - x(j)), \quad (5)$$

with $\mathcal{N}_i$ denoting the neighbours connected to vertex i by an edge.

An eigendecomposition of the Laplacian matrix, $\mathcal{L} = U\Lambda U^T$, gives a set of orthonormal eigenvectors $U = [u_0, \ldots u_{N-1}] \in \mathbb{R}^{N \times N}$ with associated real, non-negative eigenvalues $\Lambda =$

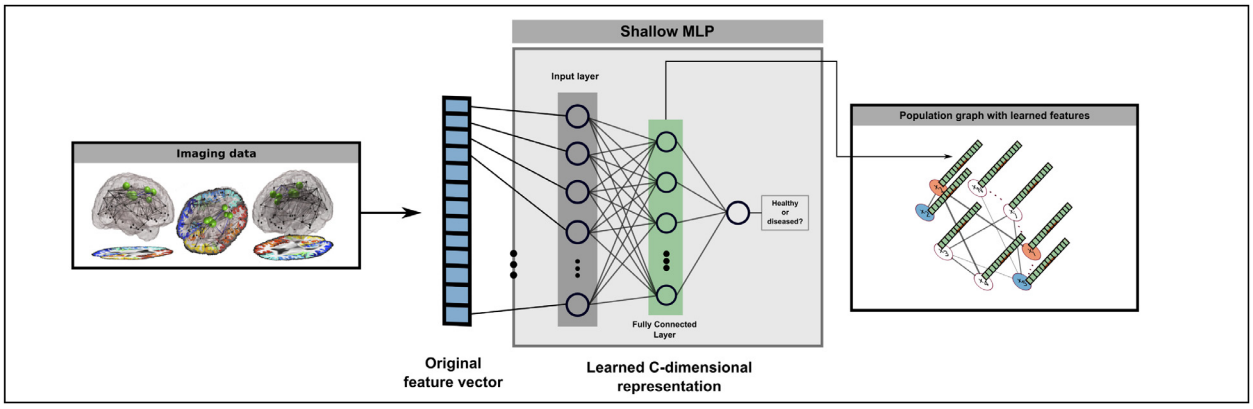


Fig. 2. Illustration of the MLP approach for ABIDE feature selection.

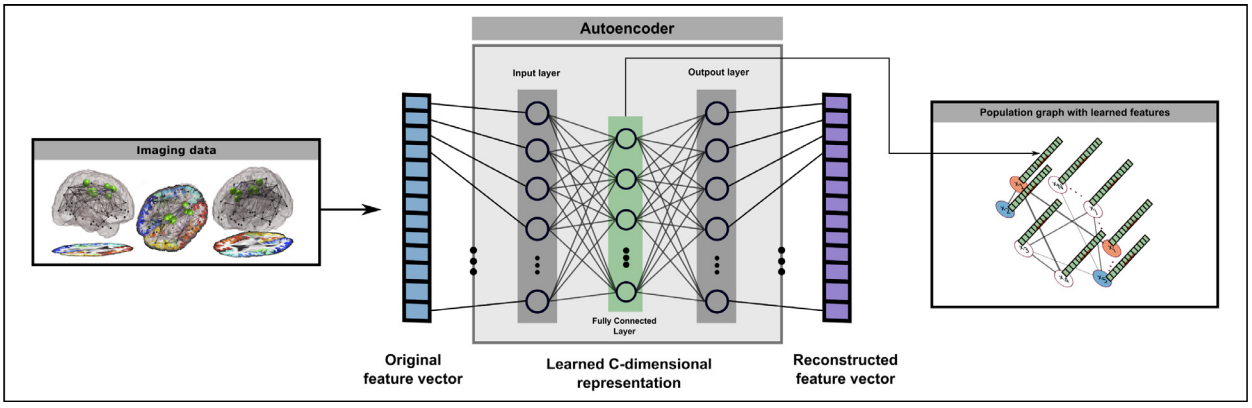


Fig. 3. Illustration of the Autoencoder approach for ABIDE feature selection.

$diag([\lambda_0, \ldots, \lambda_{N-1}]) \in \mathbb{R}^{N \times N}$. The eigenvectors associated with low frequencies/eigenvalues vary slowly across the graph, meaning that vertices connected by an edge of large weight have similar values in the corresponding locations of these eigenvectors.

Considering a spatial signal $\mathbf{x}$ defined on graph $\mathcal{G}$, its Fourier transform is defined as $\hat{\mathbf{x}} \doteq U^T\mathbf{x} \in \mathbb{R}^N$, while the inverse transform is given by $\mathbf{x} \doteq U\hat{\mathbf{x}}$. A spectral convolution of signal $\mathbf{x}$ with a filter $g_\theta = diag(\theta)$ defined in the Fourier domain can then be defined as a multiplication in the Fourier domain:

$$g_\theta * \mathbf{x} = g_\theta(\mathcal{L})\mathbf{x} = g_\theta(U\Lambda U^T)\mathbf{x} = Ug_\theta(\Lambda)U^T\mathbf{x}, \quad (6)$$

where $\theta \in \mathbb{R}^N$ are the parameters of filter g_θ.

Following the work of Defferrard et al. (2016), we restrict the class of considered filters to polynomial filters $g_\theta(\Lambda) = \sum_{k=0}^{K} \theta_k \Lambda^k$. This approach has two main advantages: 1) it yields filters that are strictly localised in space (a K-order polynomial filter is strictly K-localised) and 2) it significantly reduces the computational complexity of the convolution operator. Indeed, such filters can be well approximated by a truncated expansion in terms of Chebyshev polynomials which can be computed recursively.

It should be noted that additional simplifications on graph convolution layers are proposed in Kipf and Welling (2016) with applications to citation networks. Nonetheless, this simpler approach was not adapted to our relatively smaller and more complex datasets, which required the less efficient, yet more powerful and expressive approach, described in Defferrard et al. (2016).

2.4.2. GCN model

Our model architecture is illustrated in Fig. 4. The model is relatively simple and consists of a fully convolutional GCN with L hidden layers activated using the Rectified Linear Unit (ReLU) function. The output layer is followed by a softmax activation function. The graph is trained using the whole population graph as input. The training set comprises a labelled subset of graph nodes on which the loss function is evaluated and gradients are back propagated. The test set features (remaining unlabelled graph nodes) are observed during training, and influence the convolutions of labelled samples, making this a semi-supervised classification scheme. We further use a cross entropy loss function for the optimisation process. After training the GCN model, the softmax activations are computed on the test set, and the unlabelled nodes are assigned the labels maximising the softmax output.

3. Results

3.1. Experimental set-up

We evaluate our model on both the ADNI and ABIDE databases using a 10-fold stratified cross validation strategy. The use of 10 folds facilitates the comparison with the ABIDE state of the art (Abraham et al., 2017) where a similar strategy is adopted. To provide a fair evaluation for ADNI, we ensure that all longitudinal acquisitions of the same subject are in the same fold (i.e. either the testing or training fold). GCN parameters that are not explored in this paper are fixed and chosen according to Parisot et al. (2017), where they were optimised for the tasks considered here using a grid search. For ABIDE, we use: $L = 1$, dropout rate: 0.3, l2 regularisation: 5.10^{-4}, learning rate: 0.005, epochs: 150. The parameters for ADNI are: $L = 6$, dropout rate: 0.02, l2 regularisation: 1.10^{-5}, learning rate: 0.01, epochs: 200. Finally, graph construction variables λ for ADNI similarity and θ for quantitative phenotypic measures are $\lambda = 10$ and $\theta = 2$. In this section, we refer to the graph

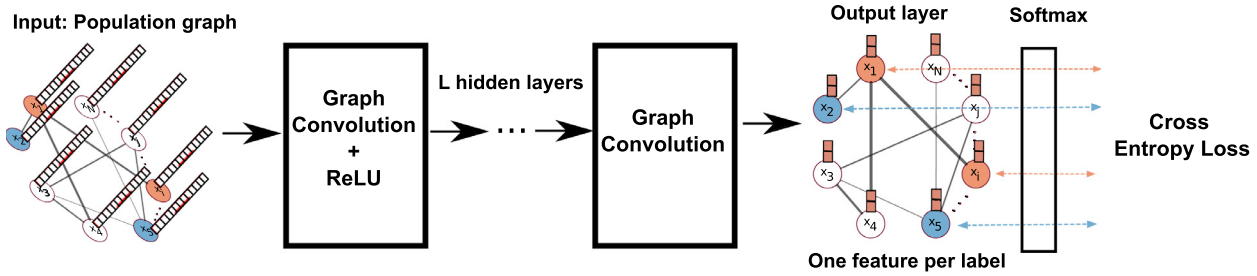


Fig. 4. Architecture of our GCN model.

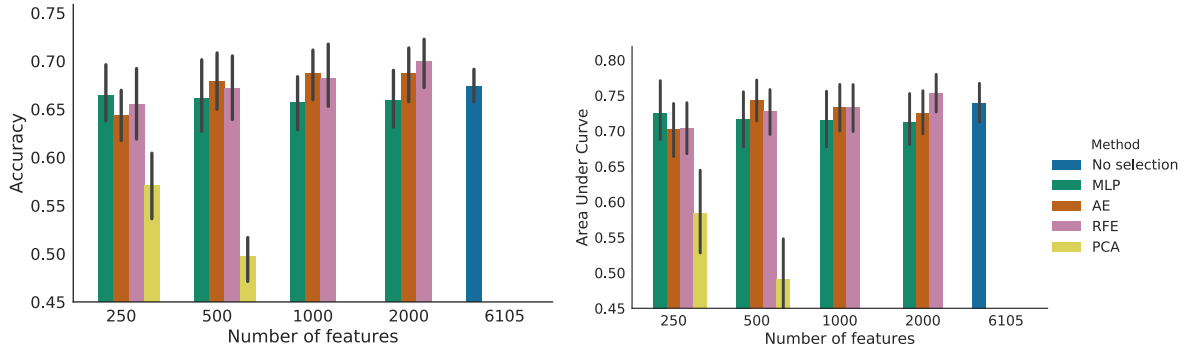


Fig. 5. ABIDE classification accuracy (a) and area under curve (b) for different feature selection strategies and number of features. The error bars report the classification accuracy across the 10 different folds.

constructed from phenotypic data, as described in Section 3.4, as the *phenotypic* graph. The default structures for experiments using the *phenotypic* graph are the ones used in Parisot et al. (2017). The phenotypic measures used to build edges are *SEX* and *AGE* for ADNI, and *SEX* and *SITE* for ABIDE. Similarly, the default polynomial order is set to $K = 3$.

3.2. ABIDE feature selection strategy

We report the influence of the feature selection scheme on the ABIDE database for the four considered methods described in Section 2.3, i.e. RFE, PCA, MLP and Autoencoder (AE). We train and evaluate a GCN using RFE, MLP and Autoencoder for $C = \{250, 500, 1000, 2000\}$ number of features. For PCA, we report results for $C = \{250, 500\}$ which correspond to approximatively 85% and 95% of explained variance. Last but not least, we report results for $C = 6105$, which corresponds to using the whole feature vector as input. AE is trained for 100 epochs with a learning rate of $5e^{-4}$, while the scikit-learn (Pedregosa et al., 2011) implementation with default parameters is used for MLP and PCA.

Results are shown in Fig. 5, reporting classification accuracy as well as Area Under Curve (AUC) for the 10 different cross validation folds. The first observation is the very poor performance of PCA as a feature selection strategy. As mentioned in Section 2.3, this approach, although very popular and efficient for the dimensionality reduction task, is limited to finding a linear projection of the features to the lower-dimensional space. Therefore, it is not as expressive as a non-linear method, like the Autoencoder. Additionally, since the ABIDE dataset is highly heterogeneous, the variance captured in the training set is likely not as representative of the test set.

MLP has the best performance with respect to other methods for 250 features but drops in performance when more features are added. This can be explained by the high tendency of the MLP classifier to overfit in our particular set up (notably due to our limited number of input samples). Features are learned from the training set, therefore optimised for classification of this specific subset of the data, which in turn, leads to overfitting when training the GCN model.

Autoencoders do not have this tendency, and our AE model has the best performance for 500 and 1000 features. RFE obtains the overall best performance for $C = 2000$ features, and its performance increases as the number of features increase. The fact that AE is performing better at lower resolution can be explained by the fact that RFE loses more information (every time the number of features is reduced, the RFE eliminates new elements of the feature vector) while AE provides a more compact representation of the whole feature vector, reducing therefore information loss with respect to RFE. The better performance of RFE for 2000 features can be explained by the fact that features are selected specifically for the classification task, while the autoencoder simply compresses the whole feature vector (which could notably exacerbate differences between sites), while at the same time being able to handle smaller database size contrarily to neural network models like the MLP. In the remainder of our experiments, we therefore use the RFE strategy with $C = 2000$ features.

3.3. Influence of polynomial order K

In a separate experiment we explored the influence of the Chebyshev polynomial order, K, which was empirically fixed to $K = 3$ in our previous work for both ABIDE and ADNI databases. Kipf and Welling (2016) tested the performance of ChebNet on three citation networks for $K \in \{1, 2, 3\}$ and found that a different degree led to optimal performance for each dataset. We believe that this difference in performance with regards to K is related to intrinsic properties of the graphs. Here, we tested $K \in \{1, 2, 3, 4, 5\}$, which corresponds to filters learned for neighbours K-hops away from the node at the centre of the receptive field.

Boxplots for mean classification performance and area under curve across 10 folds on the ABIDE database are illustrated in Fig. 6. We focus our attention to the *phenotypic* graph (using default phenotypic measures), which corresponds to the graph constructed taking into account SEX and SITE information, weighted

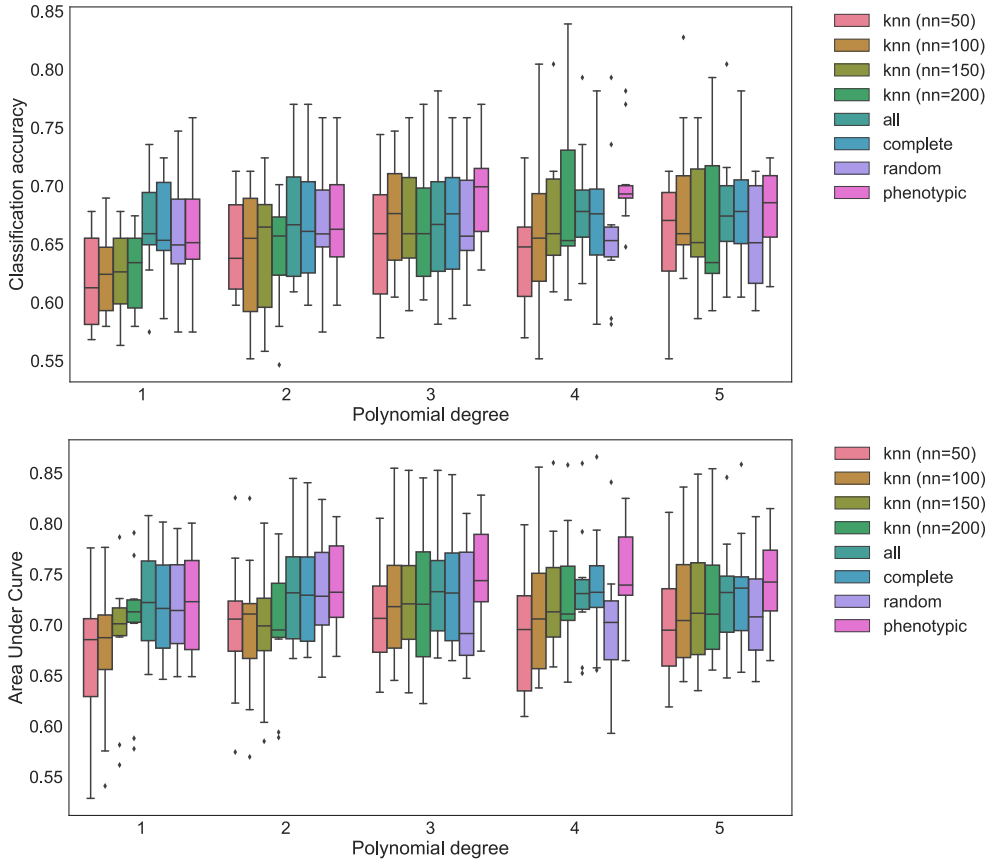


Fig. 6. Comparison of baseline graph structures and polynomial degrees for ABIDE dataset. Boxplots are generated for the performance (classification accuracy and area under curve) of each graph structure across 10 folds.

by the feature correlation similarity. This graph structure was also chosen in Parisot et al. (2017). The best average performance is achieved for $K = 4$ with accuracy 70.4% and AUC 0.75. However, this is only marginally better than the average performance achieved with $K = 3$, where accuracy reaches 69.5% and same AUC. Performance seems to increase with K, starting with mean accuracy 65.8% for $K = 1$, but then drops to 67.9% for $K = 5$.

A similar pattern is observed for the ADNI database, with results being summarised in Fig. 7. Focusing again on the *phenotypic* graph based on longitudinal information and previously used in Parisot et al. (2017), we can see that minimum performance is achieved with $K = 1$ yielding average classification 69.3% and AUC 0.76. Performance increases with K, leading to 78.3% accuracy and 0.84 AUC for $K = 3$, while for $K = 4$ mean accuracy across 10 folds is 78.8% and AUC is 0.86. Similarly to the behaviour mentioned above for the ABIDE database, performance drops to 77.7% accuracy and 0.85 AUC for $K = 5$ and decreases even further for higher values of K. All the above indicate that the receptive field is still limited for $K = 1$, while for $K = 3$ or $K = 4$ it expands enough to capture the neighbourhood structure around a node. Higher values of K likely enhance overfitting issues that arise from the limited amount of data used in comparison to tens of thousands of nodes deployed in other applications. The relationship between the optimal degree K and the diameter of the graph is yet to be explored and can lead to insightful conclusions about the degree that leads to better performance for different graph structures.

3.4. Graph construction strategy

As previously mentioned, since convolutions are parameterised on the Laplacian, the graph structure is expected to have a strong impact on semi-supervised classification performance. Therefore, we investigate the effect of different graph structures on average classification accuracy and area under curve across 10 folds. The explored graph structures include: **(a)** a k-nearest neighbours (*knn*) graph with $k = 50, 100, 150, 200$, where (dis)similarities between nodes are based solely on their associated feature information, **(b)** a binary *complete* graph, i.e. every node is connected to every other node with weight 1, **(c)** a complete graph weighted by the feature similarity between nodes (*all*), **(d)** a *random* graph with same edge density as the phenotypic graph but randomly rewired edges and **(e)** the default *phenotypic* graph structure where non-imaging information is used along with the feature similarity to construct the graph.

For the ABIDE dataset, performance for the different graph structures is summarised in Fig. 6 along with the influence of the polynomial order for each for these structures. As we can observe, the *knn* graphs are performing worse than *all, complete, random* and *phenotypic* for $K = 1$ with average accuracy 61.8% for $k = 50$ and 62.8% for $k = 200$. Additionally *all* and *complete* graphs yield equivalent results in terms of classification accuracy (66.5%), slightly outperforming the *random* (65.6%) and *phenotypic* (65.8%) graphs and indicating that for the first-order polynomial, edge density is more important than the quality of the graph. However, the *phenotypic* graph is the one leading to the best performance compared to all other graphs for $K = 3$ and $K = 4$ (as wells as best overall) followed by the *complete* graph (67.3%) for $K = 3$ and *all* graph for $K = 4$ (68.3%). Interestingly, the performance achieved with the *complete* graph is equivalent to that of a linear classifier, as presented later in Section 3.6, since all nodes contribute equally to filtering each node's features. The *phenotypic* graph is also outperforming all alternative graph structures for every K (ex-

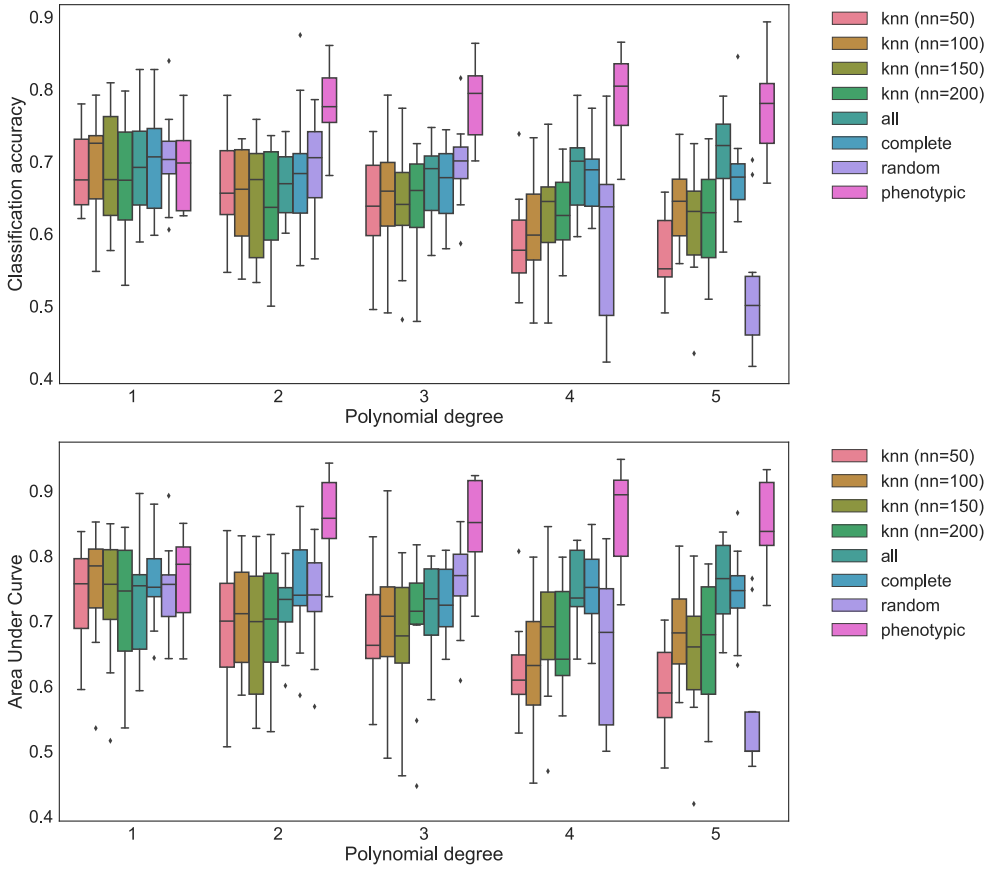


Fig. 7. Comparison of baseline graph structures and polynomial degrees for ADNI dataset. Boxplots are generated for the performance (classification accuracy and area under curve) of each graph structure across 10 folds.

cept for $K = 1$) by means of AUC, with the highest average AUC (0.75) achieved for $K = 3$. It is also worth mentioning that the *random* graph and the *knn* graphs are the ones leading to the worst performance for $K \geq 2$, since they fail to capture the complex associations/similarities between all available subjects and, thus, do not help leverage the power of graph convolutions.

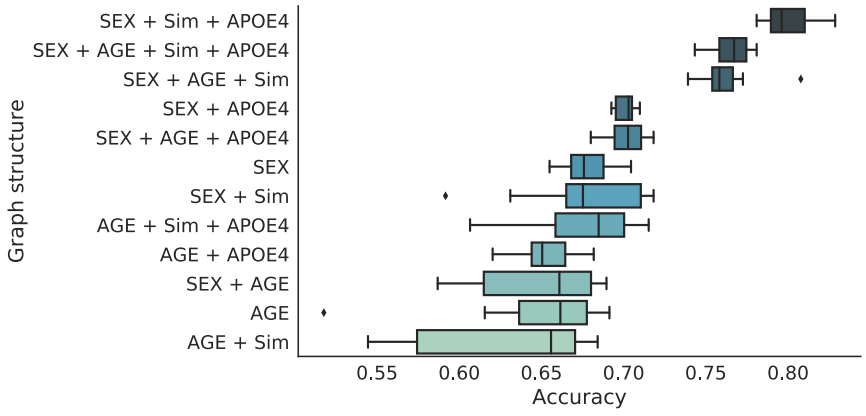
The equivalent results for the ADNI database are presented in Fig. 7. These results are slightly different from the ones presented above for the ABIDE database, since the *phenotypic* graph in this case is solely based on the longitudinal information and does not rely on the node features at all. Therefore, we observe a clearer 'superiority' of the *phenotypic* graph against all other graph structures for $K \geq 2$. Apart from that, there is a negative trend in terms of classification performance for the *knn* and *random* graphs with increasing K, demonstrating that irrelevant graphs or graphs that eliminate connections between subjects yield worse performance for high K values. The best overall mean classification accuracy is achieved with the *phenotypic* graph for $K = 4$ (78.8%) with AUC 0.86, followed by the *all* (feature-based) graph with 69.0% accuracy and 0.75 AUC. The *complete* graph is also performing slightly worse than *all* graph (accuracy 68.0%, AUC 0.75), highlighting the fact that similarity-based edge weights are meaningful and help improve the quality of label propagation.

Across the two databases, we can observe that the meaningful *phenotypic* graph leads to the best performance compared to all baseline graph structures. This pattern is more prevalent for the ADNI database, in which case the *phenotypic* graph is feature-independent. The *complete* and *all* graph structures succeed the *phenotypic* graph, especially for higher order polynomial filters, indicating that this population-based framework can benefit from interactions between all nodes, even if the structure is not optimal.

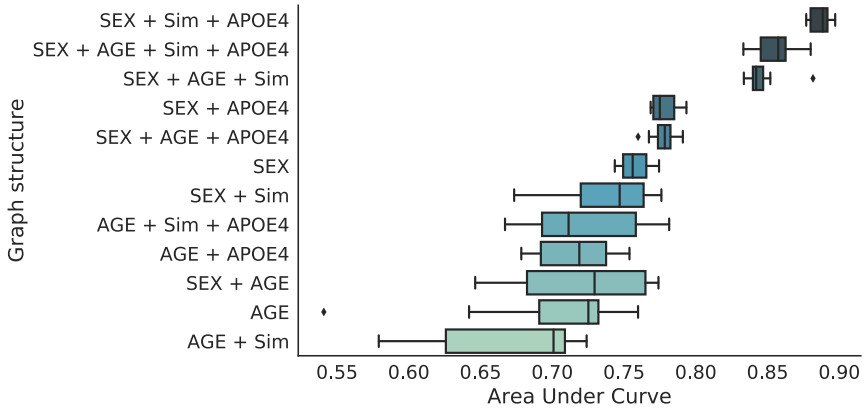
3.5. Influence of the phenotypic measures

We showed in Section 3.4 that phenotypic measures (the default measures successfully used in our previous work (Parisot et al., 2017)) led to the best graph construction when compared with alternative structures. In this section, we evaluate the performance of the model for different phenotypic graph configurations. Considering 3 different measures and a similarity function for each database, we investigate the influence of each measure on the overall classification performance by training multiple GCNs with different combinations of such measures and similarities. Results are reported for multiple initialisation seeds, so as to investigate the stability of the method with respect to the graph structure.

The default graph structure for ADNI uses the similarity link between same subjects, SEX and AGE. We, furthermore, integrate genetic information regarding the APOE4 gene. Results for multiple graph configurations using Similarity, sex, age or APOE4 measures are reported in Fig. 8. The first observation is a sharp decrease of stability across seeds (i.e. sensitivity to initialisation/lack of convergence) as the overall performance of the graph decreases. We observe a 14% difference for worse performing graph and 5% for best performing graph. This suggests that a poor graph structure significantly decreases robustness and convergence. This is to be expected as an inadequate graph structure leads to defining inaccurate neighbourhood systems, which can strongly impact the accuracy of the convolutions.



(a) Accuracy



(b) Area Under Curve

Fig. 8. Influence of the phenotypic graph structure on the classification results on the ADNI database. The boxplots report the classification accuracy across 10 different initialisation seeds.

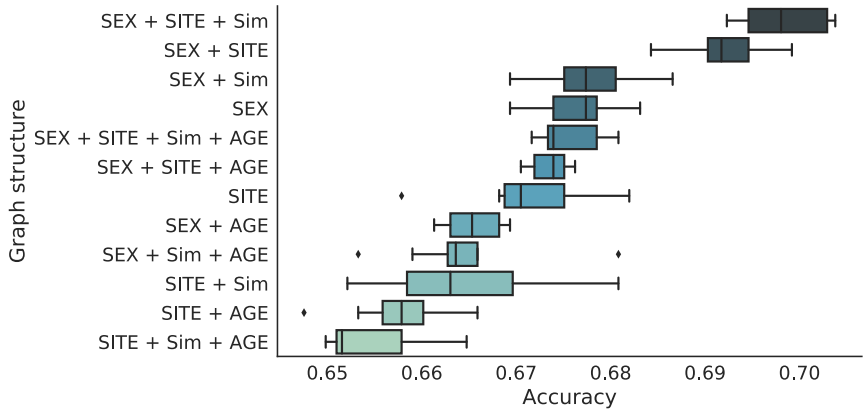
Generally, we observe that performance worsens when using AGE as a phenotypic measure, while all other measures increase performance. The similarity between same subjects significantly increases accuracy (+10% between SEX+APOE4 and SEX+APOE4+Sim), but worsens when using poor graph structures (e.g sex or age only). Last but not least, we can see that adding the APOE4 measure increases results, leading to 2 graph structures that beat the performance of the default graph. Our best performance is obtained for Similarity + SEX + APOE4, corresponding to a 80% accuracy and 0.89 AUC.

Results for the ABIDE database are reported in Fig. 9. The default ABIDE graph is based on similarity between connectivity networks, SEX and SITE. In this experiment, we also introduce the AGE parameter. Contrarily to the ADNI database, we observe very little variation between graph structures with a 3% difference in accuracy between best and worse performing graphs. Similarly, the difference between different initialisation seeds is almost negligible. The best performing graph is the one used in Parisot et al. (2017), i.e. the default graph, with an average accuracy over 10 seeds of 0.70 and 0.75. The similarity provides a small increase for almost all cases. SEX appears to provide a better accuracy than SITE, possibly due to the fact that the SITE based graph is highly disconnected. However, SITE provides better results in terms of AUC. Similarly to what was observed with ADNI, AGE consistently reduces the classification performance. This observation on both databases may be linked to the way that it is integrated in the graph, as it is the only phenotypic quantitative measure used in this paper.

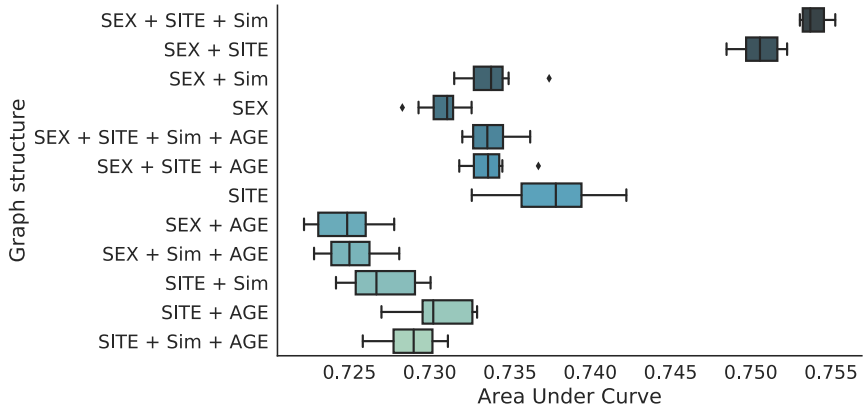
Finally, we compare two strategies to evaluate the overall performance between seeds: 1) averaging the accuracies obtained for all seeds and 2) ensembling the results using majority voting. We provide comparative results in Fig. 10 for both ADNI and ABIDE database. We can see that there is very little difference between both strategies for the ABIDE database and the best performing graphs of the ADNI database, while a significant increase in performance is obtained using ensembling of poor quality graphs which tend to have very unstable performances across seeds. The best performance is almost consistently obtained by ensembling. This suggests that the optimal strategy is to ensemble between multiple initialisations. However, the small variability between initialisations for good graph structures suggests that results are consistent across seeds and that one initialisation provides sufficient evaluation.

3.6. Comparison to other methods

Apart from investigating how the different components of the proposed framework impact semi-supervised classification performance, we further compared the GCN results to different well-established classifiers. These include a ridge classifier (using the scikit-learn library implementation (Pedregosa et al., 2011)), which showed the best performance amongst linear classifiers, a random forest classifier with 100 estimators and a multi-layer perceptron (MLP) classifier. We also add the performance reported in Parisot et al. (2017) using the GCN model (GCN) alongside the



(a) Accuracy



(b) Area Under Curve

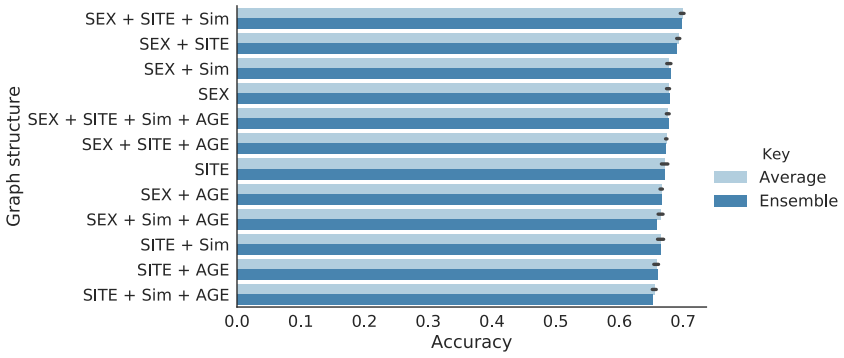
Fig. 9. Influence of the phenotypic graph structure on the classification results on the ABIDE database. The boxplots report the classification accuracy across 10 different initialisation seeds.

best performance obtained in this manuscript throughout our experiments (GCN (best)). This is to highlight the performance increase obtained in this paper, and for consistency purposes with Parisot et al. (2017). For the MLP we used the same parameters as the GCN implementation, in terms of number of hidden layers, number of features, dropout, seed, learning rate and regularisation, and fixed the number of epochs to 200 for both datasets. Comparative boxplots across all folds between these three approaches, along with the previously used model and the best achieved result (in terms of accuracy) for the same seed through our extensive evaluation in this paper are shown in Fig. 11 for both databases.

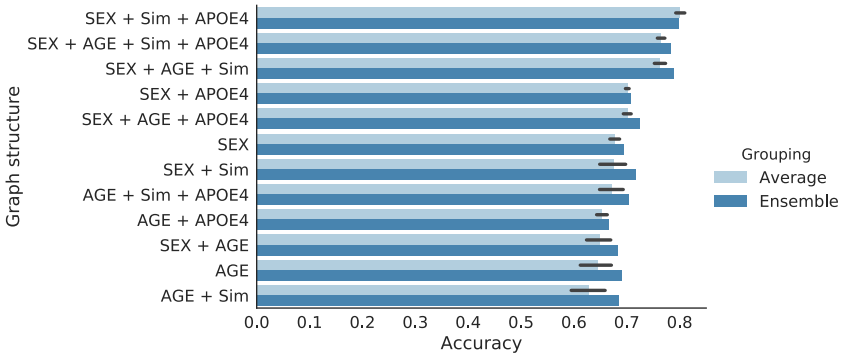
The worst performance on the ABIDE database is observed for the random forest classifier with 61.0% average accuracy and 0.64 AUC. The ridge classifier is doing better than the MLP in terms of classification accuracy (65.3% vs 63.0%), but is outperformed by the MLP in terms of AUC (0.69 vs 0.71). The best overall performance observed with the GCN model is 70.4% accuracy and 0.75 AUC, outperforming the recent state of the art (66.8%) (Abraham et al., 2017). For the ADNI database, the worst performance is achieved with the ridge classifier (67.2% accuracy and 0.74 AUC), followed by the random forest classifier (68.8% and 0.74 AUC). The MLP classifier yields slightly improved results compared to these two classifiers with 70.4% mean classification accuracy and 0.77 AUC. GCN results for this database show a large increase in performance with respect to the competing methods, with an average accuracy of 80.0% and AUC of 0.88, higher than state of the art results (Tong et al., 2017a), corresponding to a 9% increase over an MLP. Finally, regarding the GCN model, we observe a 4% increase in accuracy for the ADNI database using a better graph structure (integrating APOE4 gene information and eliminating AGE information) with respect to Parisot et al. (2017). The ABIDE performance is stable, suggesting an optimal graph structure in Parisot et al. (2017) or a limitation inherent to the dataset.

4. Discussion

In this paper, we proposed a method for group-level population diagnosis that exploits the novel concept of spectral graph convolutions. We modelled populations as a sparse graph combining subject-specific imaging data and pairwise interactions described using phenotypic and other non-imaging information. This sparse graph is used to train a GCN in a semi-supervised manner for node classification, learning on a subset of labelled nodes and evaluating on the rest. Our experiments on two large and challenging databases (ABIDE and ADNI) confirm our initial hypothesis about the importance of contextual pairwise information for classification, as we obtain state of the art performance with 70.4% (ABIDE) and 80% (ADNI) accuracy, corresponding to increases of 5% and 9% with respect to classifiers using node features only. Our extensive evaluation analyses the different components of the model, including the feature selection method, polynomial degree and graph construction strategy. Exploring different graph structures and baselines, we show how our phenotypic graph formulation yields more accurate and stable results, as well as the impor-

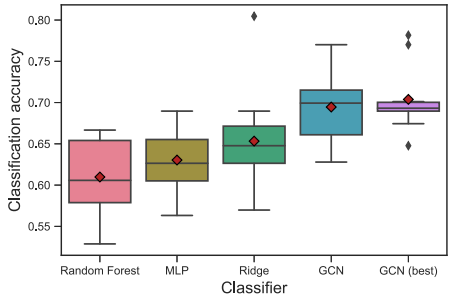


(a) ABIDE

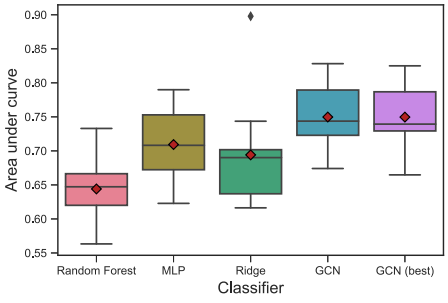


(b) ADNI

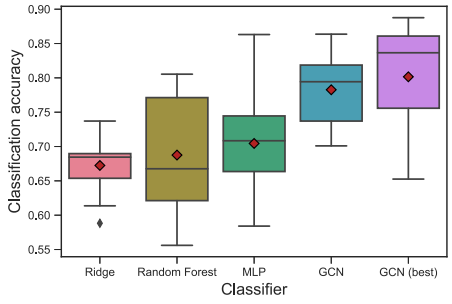
Fig. 10. Classification accuracy for the ABIDE and ADNI database. Results are reported across 10 different initialisation seeds, either by averaging results across seeds (light blue), or by ensembling all results (dark blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)



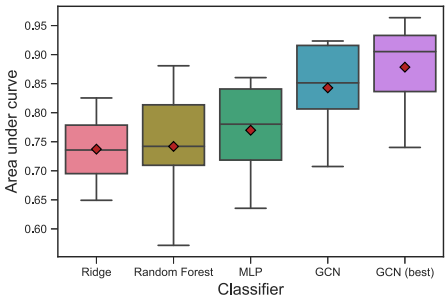
(a) ABIDE accuracy



(b) ABIDE AUC



(c) ADNI accuracy



(d) ADNI AUC

Fig. 11. Comparative boxplots of the classification accuracy and area under curve (AUC) over all cross validation folds for the (a, b) ABIDE and (c, d) ADNI databases (MCI conversion task) for different baseline classifiers (random forest, ridge and MLP), the GCN model used in Parisot et al. (2017) and the best GCN model from the current extensive evaluation.

tance of choosing appropriate phenotypic measures to model the pairwise interactions.

While our method is tested against multiple baselines, conditional random fields (CRF)/ Markov random fields (MRF) models also show interesting similarities with our approach. As in our setting, CRF models are cast as graph labelling problems and seek to increase classification performance by modelling pairwise interactions between graph nodes. Nonetheless, such approaches require substantial methodological decisions (definitions of node likelihood from a separate learning strategy, devising an optimisation strategy adapted to the considered graph structure and optimising parameters), making direct comparison not straightforward. Furthermore, CRF formulations model feature vectors and pairwise interactions sequentially, leading to a weaker model than our approach which learn end-to-end features representation informed by pairwise interactions.

In this paper, we cast the problem of AD and ASD diagnosis as a binary classification due to the annotations provided in the complete ABIDE and ADNI databases. However, it is well known that both diseases lie on a spectrum (Mega et al., 1996; Volkmar et al., 2009). One could therefore be seeking to carry out multi-class classification or to predict continuous outputs in such setting. This can be done easily using our framework by updating the output size and loss function to predict multi-class labels or carry out a regression task.

The ABIDE database is particularly challenging due to the fact that images are acquired at different sites with different protocols. This strongly reduces the comparability of image features from different sites. Our experiments suggest that an accuracy of approx. 70% could be an intrinsic limitation of the whole dataset. This observation aligns with the most recent works on the ABIDE dataset, reporting 68% (Abraham et al., 2017) and 70% (Heinsfeld et al., 2018) accuracies on the whole dataset (using leave one out cross validation for the latter). It is likely that increasing performance on this dataset would require models that can learn to eliminate this inter-site variability by capturing site-invariant class-discriminative patterns. It should be noted that significantly better accuracy results can be obtained when restricting analysis on one site only or when integrating cognitive test results.

Spectral GCNs provide a powerful and principled way of performing convolutions on irregular graph structures. In this paper, we use Chebyshev polynomials, which have proven to be an excellent compromise between speed and accuracy. The recently introduced Cayley polynomials (Levie et al., 2017) have shown a lot of potential and could provide further improvements of our results. Despite their many advantages, spectral GCNs have an important drawback, which is that they can only be applied to graphs of fixed structure due to their parametrisation on the graph Laplacian. While a small modification of the graph structure (e.g. replacing a node) is unlikely to alter performance, retraining the GCN will be a necessity for substantially modified graphs. This also poses a challenge when new subjects become available and need to be incorporated in the analysis, as the model will need to be trained from scratch. In situations where one expects highly variable graph structures, spatial GCNs (Monti et al., 2016) should be considered.

Another limitation is the use of hand crafted features as node descriptors. The main advantage of deep learning is that the network learns itself an optimal feature representation of the raw input data for the task at hand. Ideally, one would want an end to end strategy, learning optimal node features from the input images or connectivity networks. This could potentially be achieved using the method proposed in Monti et al. (2017) for matrix completion, proposing a multi-graph convolution method. In our case, the second graph could be the input image data or connectivity networks. Despite their limited performance in our experiments, feature encoding using autoencoders or MLPs have the potential of providing better results than RFE. Here we use very simple models for both AE and MLP; therefore, their performance is likely to increase with appropriate parameter and model engineering. Nonetheless, the main limitation of those neural network based approaches compared to RFE is the limited amount of training data, especially for the smallest sites. Using additional data (for example using the recently released ABIDE II dataset) has the potential to increase the performance of such methods, and could also allow using adversarial techniques to learn a feature representation that is independent of the acquisition site. Last but not least, the weak performance of the considered learning based methods (AE, MLP, PCA) could be due to an excessing amount of data learning. Learning a representation from existing relationships within the dataset, and repeating the same process in the GCN network, amounts to double learning the data which could reduce performance. On the contrary, RFE does not learn a new representation of the feature vector.

Devising an effective strategy to construct the population graph is essential and far from obvious. We have explored several graph structures in this paper and demonstrated how the graph can significantly affect classification accuracy. Our phenotypic graph construction strategy yielded the best performance, nonetheless, each graph edge comprises multiple types of information. Our experiments have shown that integrating the wrong or redundant phenotypic information (e.g. patient age in our case) has a strong negative impact on the results, while accurate measures with known links to the pathologies substantially increase performance. This is in accordance with our initial hypothesis that integrating non-imaging data that is known to influence the imaging data or subject's label yields better feature representations.

Another graph construction strategy could be a pure learning based approach, learning a graph structure from self-attention weights and using all potential phenotypic measures as features. An added value of such an approach could be the identification of new important phenotypic measures by exploration of learned attention weights. We could also consider attributed graph edges (i.e. a vector instead of a scalar weight) comprising all measures. Such a model could be exploited using a recent spatial GCN strategy (Simonovsky and Komodakis, 2017). The proposed framework could benefit from more sophisticated graph representation learning techniques, which can help the discovery of edge features and contribute to improved transfer-learning and node classification results (Rossi et al., 2017). One could also build multiple graph structures (e.g. one per phenotypic measure, or a mixture of phenotypic and knn graphs) and find a common Fourier base for convolutions via joint diagonalisation of the graphs' Laplacian matrices (Eynard et al., 2015).

Among the main limitations of this work one should consider the generalisation of this framework to unseen sites, e.g. in the ABIDE case. Since this is an application of transductive learning, generalisation to new unseen domains is expected to lead to performance decrease, especially if the training dataset is not large enough to capture population variability. Moreover, the way the ADNI graph is constructed, with multiple scans per subject being modelled as nodes and classified independently, is likely to introduce biases towards subjects with more visits available. Last but not least, highly class imbalanced problems constitute a scenario that would require further research. In this work, we performed two studies with relatively balanced data. However, in certain types of population studies (e.g. genome-wide predictions tasks (Yones et al., 2017)) one can find huge class imbalance ratios, in the order of 1:10000. In the future, we would like to study how graph convolutions can be used to leverage the available annotated data towards improving prediction rate in highly class imbalanced problems.

Acknowledgements

This work was supported by the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement no. 319456. The Titan X Pascal used for this research was donated by the NVIDIA Corporation. Enzo Ferrante is beneficiary of an AXA Research grant. Sofia Ira Ktena is supported by the EPSRC Centre for Doctoral Training in High Performance Embedded and Distributed Systems (HiPEDS, Grant Reference EP/L016796/1).

References

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. NeuroImage 147, 736–745.

Anirudh, R., Thiagarajan, J. J., 2017. Bootstrapping graph convolutional neural networks for autism spectrum disorder classification. arXiv preprintarXiv:1704.07487.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145, 137–165.

Bronstein, M.M., Bruna, J., LeCun, Y., Szlam, A., Vandergheynst, P., 2017. Geometric deep learning: going beyond euclidean data. IEEE Signal Process. Mag. 34 (4), 18–42.

Brosch, T., Tam, R., 2013. Manifold learning of brain MRIs by deep learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 633–640. doi:10.1007/978-3-642-40763-5_78.

Bruna, J., Zaremba, W., Szlam, A., LeCun, Y., 2013. Spectral networks and locally connected networks on graphs.. CoRR abs/1312.6203.

Craddock, C., Sikka, S., Cheung, B., Khanuja, R., Ghosh, S.S., Yan, C., Li, Q., Lurie, D., Vogelstein, J., Burns, R., et al., 2013. Towards automated analysis of connectomes: the configurable pipeline for the analysis of connectomes (c-pac). Front. Neuroinform. 42.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.-O., Chupin, M., Benali, H., Colliot, O., Initiative, A.D.N., et al., 2011. Automatic classification of patients with Alzheimer's disease from structural mri: a comparison of ten methods using the adni database. Neuroimage 56 (2), 766–781.

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. In: NIPS, pp. 3837–3845.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. Neuroimage 31 (3), 968–980.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19 (6), 659.

Eynard, D., Kovnatsky, A., Bronstein, M.M., Glashoff, K., Bronstein, A.M., 2015. Multimodal manifold analysis by simultaneous diagonalization of laplacians. IEEE Trans. Pattern Anal. Mach. Intell. 37 (12), 2505–2517.

Guo, Y., Nejati, H., Cheung, N.-M., 2017. Deep neural networks on graph signals for brain imaging analysis. arXiv preprintarXiv:1705.04828.

Hammond, D.K., Vandergheynst, P., Gribonval, R., 2011. Wavelets on graphs via spectral graph theory. Appl. Comput. Harmon. Anal. 30 (2), 129–150.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31.

Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F., 2018. Identification of autism spectrum disorder using deep learning and the abide dataset. NeuroImage 17, 16–23.

Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H., 2016. Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2424–2433.

Kana, R.K., Uddin, L.Q., Kenet, T., Chugani, D., Müller, R.-A., 2014. Brain connectivity in autism. Front. Hum. Neurosci. 8.

Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. Brainnetcnn: convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage 146, 1038–1049.

Kipf, T. N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprintarXiv:1609.02907.

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. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 469–477.

Ktena, S.I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., Rueckert, D., 2018. Metric learning with spectral graph convolutions on brain connectivity networks. NeuroImage 169, 431–442.

Ledig, C., Heckemann, R.A., Hammers, A., Lopez, J.C., Newcombe, V.F., Makropoulos, A., Lötjönen, J., Menon, D.K., Rueckert, D., 2015. Robust whole-brain segmentation: application to traumatic brain injury. Med. Image Anal. 21 (1), 40–58.

Levie, R., Monti, F., Bresson, X., Bronstein, M. M., 2017. Cayleynets: graph convolutional neural networks with complex rational spectral filters. arXiv preprintarXiv:1705.07664.

Lombaert, H., Criminisi, A., Ayache, N., 2015. Spectral forests: learning of surface data, application to cortical parcellation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 547–555.

Masci, J., Boscaini, D., Bronstein, M., Vandergheynst, P., 2015. Geodesic convolutional neural networks on Riemannian manifolds. In: Proceedings of the IEEE international conference on computer vision workshops, pp. 37–45.

Mega, M.S., Cummings, J.L., Fiorello, T., Gornbein, J., 1996. The spectrum of behavioral changes in Alzheimer's disease. Neurology 46 (1), 130–135.

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. arXiv preprintarXiv:1611.08402.

Monti, F., Bronstein, M. M., Bresson, X., 2017. Geometric matrix completion with recurrent multi-graph neural networks. arXiv preprintarXiv:1704.06803.

Mosconi, L., Perani, D., Sorbi, S., Herholz, K., Nacmias, B., Holthoff, V., Salmon, E., Baron, J.-C., De Cristofaro, M., Padovani, A., et al., 2004. Mci conversion to dementia and the apoe genotype a prediction study with fdg-pet. Neurology 63 (12), 2332–2340.

Niepert, M., Ahmed, M., Kutzkov, K., 2016. Learning convolutional neural networks for graphs. arXiv preprintarXiv:1605.05273.

Parisot, S., Darlix, A., Baumann, C., Zouaoui, S., Yordanova, Y., Blonski, M., Rigau, V., Chemouny, S., Taillandier, L., Bauchet, L., et al., 2016. A probabilistic atlas of diffuse who grade ii glioma locations in the brain. PLoS ONE 11 (1), e0144200.

Parisot, S., Ktena, S.I., Ferrante, E., Lee, M.C.H., Moreno, R.G., Glocker, B., Rueckert, D., 2017. Spectral graph convolutions for population-based disease prediction. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 177–185.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12 (Oct), 2825–2830.

Ries, M.L., Carlsson, C.M., Rowley, H.A., Sager, M.A., Gleason, C.E., Asthana, S., Johnson, S.C., 2008. Magnetic resonance imaging characterization of brain structure and function in mild cognitive impairment: a review. J. Am. Geriatr. Soc. 56 (5), 920–934.

Rossi, R. A., Zhou, R., Ahmed, N. K., 2017. Deep feature learning for graphs. arXiv preprintarXiv:1704.08829.

Rudie, J.D., Brown, J., Beck-Pancer, D., Hernandez, L., Dennis, E., Thompson, P., Bookheimer, S., Dapretto, M., 2013. Altered functional and structural brain network organization in autism. NeuroImage 2, 79–94.

Sahiner, B., Chan, H.-P., Petrick, N., Wei, D., Helvie, M.A., Adler, D.D., Goodsitt, M.M., 1996. Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans. Med. Imaging 15 (5), 598–610.

Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G., 2009. The graph neural network model. IEEE Trans. Neural Netw. 20 (1), 61–80.

Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P., 2013. The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process. Mag. 30 (3), 83–98.

Simonovsky, M., Komodakis, N., 2017. Dynamic edge-conditioned filters in convolutional neural networks on graphs. arXiv preprintarXiv:1704.02901.

Soper, H.V., Satz, P., Orsini, D.L., Henry, R.R., Zvi, J.C., Schulman, M., 1986. Handedness patterns in autism suggest subtypes. J. Autism Dev. Disord. 16 (2), 155–167.

Thompson, P.M., Stein, J.L., Medland, S.E., Hibar, D.P., Vasquez, A.A., Renteria, M.E., Toro, R., Jahanshad, N., Schumann, G., Franke, B., et al., 2014. The enigma consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain Imaging Behav. 8 (2), 153–182.

Tong, T., Gao, Q., Guerrero, R., Ledig, C., Chen, L., Rueckert, D., 2017. A novel grading biomarker for the prediction of conversion from mild cognitive impairment to Alzheimer's disease.. IEEE Trans. Biomed. Eng. 64 (1), 155–165. doi:10.1109/TBME.2016.2549363.

Tong, T., Gray, K., Gao, Q., Chen, L., Rueckert, D., Initiative, A.D.N., et al., 2017. Multi–modal classification of Alzheimer's disease using nonlinear graph fusion. Pattern Recognit. 63, 171–181.

Volkmar, F.R., State, M., Klin, A., 2009. Autism and autism spectrum disorders: diagnostic issues for the coming decade. J. Child Psychol. Psychiatr. 50 (1–2), 108–115.

Wang, Z., Zhu, X., Adeli, E., Zhu, Y., Nie, F., Munsell, B., Wu, G., 2017. Multi-modal classification of neurodegenerative disease by progressive graph-based transductive learning. Med. Image Anal. 39, 218–230.

Werling, D.M., Geschwind, D.H., 2013. Sex differences in autism spectrum disorders.. Curr. Opin. Neurol. 26 (2), 146–153.

Wolz, R., Aljabar, P., Hajnal, J.V., Lötjönen, J., Rueckert, D., 2012. Nonlinear dimensionality reduction combining mr imaging with non-imaging information. Med. Image Anal. 16 (4), 819–830.

Yones, C., Stegmayer, G., Milone, D.H., 2017. Genome-wide pre-miRNA discovery from few labeled examples. Bioinformatics.

Zhao, M., Chan, R.H., Chow, T.W., Tang, P., 2014. Compact graph based semi-supervised learning for medical diagnosis in alzheimer's disease. IEEE Signal Process. Lett. 21 (10), 1192–1196.