A Comparative Study of Graph Neural Networks for Shape Classification in Neuroimaging

Nairouz Shehata N.MOHAMED16@IMPERIAL.AC.UK
Wulfie Bain WULFIE.BAIN@OUTLOOK.COM
Ben Glocker B.GLOCKER@IMPERIAL.AC.UK
Department of Computing, Imperial College London, UK

Editors: Jelmer Wolterink, Angelica I. Aviles-Rivero, Erik Bekkers

Abstract

Graph neural networks have emerged as a promising approach for the analysis of non-Euclidean data such as meshes. In medical imaging, mesh-like data plays an important role for modelling anatomical structures, and shape classification can be used in computer aided diagnosis and disease detection. However, with a plethora of options, the best architectural choices for medical shape analysis using GNNs remain unclear.

We conduct a comparative analysis to provide practitioners with an overview of the current state-of-the-art in geometric deep learning for shape classification in neuroimaging. Using biological sex classification as a proof-of-concept task, we find that using FPFH as node features substantially improves GNN performance and generalisation to out-of-distribution data; we compare the performance of three alternative convolutional layers; and we reinforce the importance of data augmentation for graph based learning. We then confirm these results hold for a clinically relevant task, using the classification of Alzheimer's disease.

Keywords: Shape classification, graph neural networks, brain structures, 3D mesh data

1. Introduction

Geometric deep learning generalizes classical neural network models to non-Euclidean domains such as point clouds, graphs, or meshes (Wu et al., 2020). It has therefore become popular across various fields from computer vision (Zhou et al., 2020b) and physics (Shlomi et al., 2020), to healthcare topics (Dash et al., 2019) such as disease prediction (Kazi et al., 2019), drug discovery (Li et al., 2017), and brain connectome analysis (Kim et al., 2021).

A recent study (Sarasua et al., 2022) investigated the expressiveness of mesh representations for disease classification. We complement these findings by conducting a comparative study evaluating different graph neural networks (GNNs) for the classification of anatomical meshes extracted from neuroimaging data. We propose a simple yet effective multi-graph architecture with a shared submodel for learning shape embeddings (see Fig. 1). Different graph convolutional layers are compared; GCNConv (Kipf and Welling, 2016), GraphConv (Morris et al., 2019), and SplineCNN (Fey et al., 2018). In all cases, we observe substantial performance improvements when using Fast Point Feature Histograms (FPFH) as node features, which to our knowledge has not been explored before. We also investigate the effect of data augmentation, finding improvements in generalization to data from new domains. Our findings on the proof-of-concept task of biological sex classification are confirmed on the clinically relevant diagnostic task of Alzheimer's disease classification.

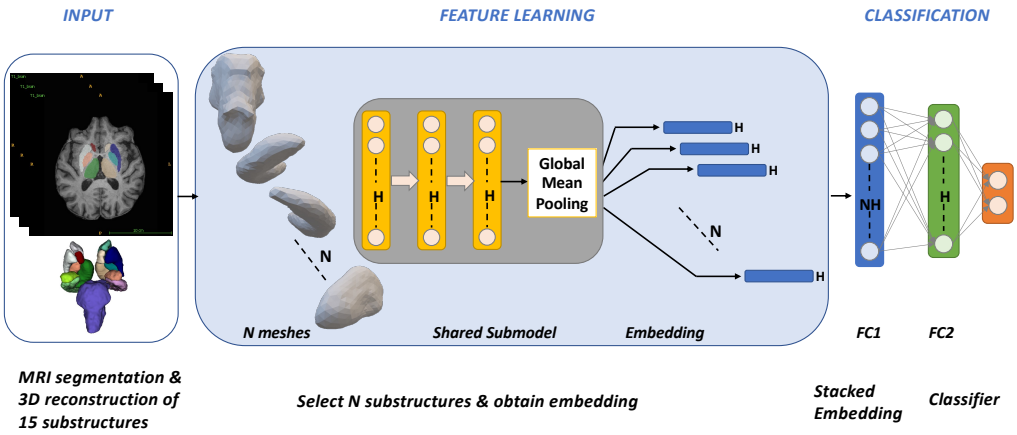


Figure 1: Proposed multi-graph architecture; N is the number of meshes (here, N=15), H is the number of hidden features (H = 32), and FC is a fully connected layer.

2. Graph Neural Network Architecture

As the field of geometric deep learning has expanded, the architectural choices available to practitioners has proliferated. Here we outline our approach on three key aspects: the type of convolutional layer used, the number of convolutional submodels, and the type of geometric features encoded at the node level.

2.1. Selected Graph Convolutional Operators

Graph convolutional operations are analogous to CNN operations on images, respecting the additional invariants that arise in this domain, permutation invariance being key due to the artificial ordering of nodes that arises when representing graphs. As shown in Bronstein et al. (2021), many GNNs follow a blueprint of 'message passing' (Gilmer et al., 2017), whereby node features are updated using an aggregation on the features of nodes in their neighbourhood, but there is significant variance in how this is done. In this paper, we compare three seminal graph convolutional layers from the literature: GCNConv (Kipf and Welling, 2016), GraphConv (Morris et al., 2019), and SplineCNN (Fey et al., 2018). These are selected as popular representatives of graph convolutional layers, that are easy to use as plugin replacements in generic architectures. We direct readers to the original papers for details. Existing literature has compared GCNConv and GraphConv (Xu et al., 2018; Morris et al., 2019), and we extend this to medical imaging.

2.2. Multi-graph Architecture

As multiple subcortical structure subgraphs may be extracted simultaneously from a single sample brain scan, one must also choose how to utilise these. One option is to combine them

into a single multigraph per sample (Wang et al., 2021; Chaari et al., 2022). However, it might not be obvious how to define edges between graphs of different anatomical structures. Alternatively, as in this paper, each subgraph can be input to a specific GNN, and the results combined into a sample level output. Practitioners must decide the number of GNNs to use. One approach is a single shared GNN that learns from all subgraphs, while another is inputting each subgraph to a separate GNN i.e. the number of (sub) GNNs is equal to the number of subgraphs per sample (Hong et al., 2021). The latter approach allows each (sub) GNN to learn structure specific embeddings, whilst the former encourages the GNN to generalise learnings across structures.

Initially, we tested both a single shared and structure specific GNN submodel, finding that the performance was comparable. Using a shared submodel significantly reduces the number of parameters. Given considerations on neural networks training time (Li, 2020), cost (Wiggers, 2020), and environmental impact (Strubell et al., 2019), our preliminary results led us to use a shared GNN in this paper: each brain substructure is passed to the shared submodel to obtain an embedding. We use three convolutional layers in the submodel with ReLU activations. A global average pooling layer is used as a readout layer to aggregate the node representations into one graph embedding. These embeddings are then stacked and passed through a fully connected layer for final classification (cf. Fig. 1).

2.3. Node and edge representation

The meshes representing anatomical brain structures are defined by a set of nodes and edges, where both can carry additional information. Nodes can encode arbitrary feature vectors, from spatial information such as mesh coordinates to more complex, geometric feature descriptors. In computer vision, hand crafted features based on carefully designed descriptors have been largely abandoned in the end-to-end deep learning paradigm (Battaglia et al., 2018). However, in the case of shape analysis, we believe there is value in sophisticated, geometrical feature extractors, especially when there are limited amounts of training data. We evaluate the use Fast Point Feature Histograms (FPFH) (Rusu et al., 2009) as node features, and compare these with positional node features in form of Cartesian coordinates, and no node features (realized by setting constant values).

To calculate the FPFH features on a mesh, first a point feature histogram is computed: for each query point p_r, all neighbouring points inside a 3D sphere of radius r centered at point p_r are selected (k-neighbourhood points); then, for each pair p_r and p_k in the k-neighbourhood points of p_r, their normals are estimated as n_r and n_k. The point with the smaller angle between the line joining the pair of points and the estimated normals is chosen to be p_r. Finally a *Darboux frame* is defined as $(u = n_r, v = (p_k - p_r)u, w = u \times v)$ and the angular variations of n_r and n_k are computed:

$$\alpha = v \cdot n_k$$

$$\phi = (u \cdot (p_k - p_r))/ \left\|(p_k - p_r)\right\|$$

$$\theta = arctan(w \cdot n_k, u \cdot n_k)$$

Second, a Simple Point Feature Histogram (SPFH) is obtained by calculating the point features of each neighboring point p_k (Rusu et al., 2008). Finally, to calculate FPFH,

the SPFH of the k neighbours are used to calculate the final histogram of p_r, where they are weighted by the distances between p_r and the neighbours p_k. N is the number of points within the sampling radius (number of neighbours to the reference point). In our implementation, the sampling radius was set to 10mm and maximum number of neighbours to 100.

$$FPFH(p_r) = SPFH(p_r) + \frac{1}{N}\sum_{k=1}^{N}\frac{SPFH(p_k)}{\|p_k - p_r\|}$$

Besides the node features, we also encode edge attributes in terms of relative spherical coordinates between two nodes. Edge attributes are processed only within SplineCNN layers, but otherwise ignored in both GCNConv and GraphConv layers, as these can only use edge weights and not attributes.

3. Datasets

We utilize four neuroimaging datasets to test generalization and robustness of the classification performance. We use data from the UK Biobank imaging study (UKBB)[1] (Sudlow et al., 2015; Miller et al., 2016), the Cambridge Centre for Ageing and Neuroscience study (Cam-CAN) (Shafto et al., 2014; Taylor et al., 2017), and the IXI dataset[2]. Both UKBB and Cam-CAN use a similar imaging protocol with Siemens 3T scanners. IXI consists of data acquired at three different sites including Guy's Hospital using a Philips 1.5T system, Hammersmith Hospital using a Philips 3T scanner, and Institute of Psychiatry using a GE 1.5T system. UKBB, Cam-CAN, and IXI are data from healthy volunteers. We only discarded data related to subjects whose sex or age entries were unavailable.

We also use the OASIS-3 dataset with 716 cognitively normal participants and 318 participants who reach various stages of cognitive decline during the study, allowing Alzheimer's disease (AD) related tasks such as classification (LaMontagne et al., 2019). The cognitive status is reflected in the clinical dementia rating (CDR) that accompanies the imaging dataset, with subjects receiving a score of: 0 for normal, 0.5 for very mild dementia, 1 for mild dementia, 2 for moderate dementia and 3 for severe dementia (Morris, 1991). The CDR is collected in clinical sessions, separate to the imaging sessions, meaning sessions must be 'matched' to get an {image, CDR score} pair. We match the clinical diagnosis closest in time to each scan, before filtering out samples where the absolute time difference between scan and clinical assessment is greater than 365 days. To avoid difficulties in assigning scans to training, validation, and testing, we only use one scan per subject, leaving 1,084 unique scans. We exclude 50 samples because their sex or age information was missing. The final set of 1,034 comprises 716, 188, 111, 18 and 1 samples, for CDR of 0, 0.5, 1, 2 and 3 respectively. We binarize CDR to 0 and 1 (for CDR score 0.5, 1, 2 and 3).

The UKBB data comes pre-processed with already extracted meshes for 15 subcortical brain structures[3]. We apply our own processing pipeline to Cam-CAN, IXI and OASIS-3 to match UKBB as closely as possible: 1) Skull stripping with ROBEX v1.2[4] (Iglesias et al.,

1. UK Biobank Resource under Application Number 12579
2. https://brain-development.org/ixi-dataset/
3. Brain stem, left/right thalamus, caudate, putamen, pallidum, hippocampus, amygdala, accumbens-area
4. https://www.nitrc.org/projects/robex

2011); 2) Bias field correction with N4ITK[5] (Tustison et al., 2010); 3) Sub-cortical brain structure segmentation and meshing using FSL FIRST [6] (Patenaude et al., 2011).

Table 1: Number of samples, percentage of females, and mean, min, and max age.

Dataset	Samples	Female (%)	Age (years)
UKBB	13,749	47	61 [44, 73]
Cam-CAN	652	51	54 [18, 88]
IXI	563	58	49 [20, 86]
OASIS-3	1,034	55	72 [42, 97]

4. Experiments

The experiments were designed to evaluate and compare three main aspects: (i) the choice of convolutional layers for the shared submodel; (ii) the choice for the node features; (iii) the effect of data augmentation on robustness and generalization.

4.1. Implementation and Training

We use the Adam optimizer with a learning rate of 0.001 and the standard cross entropy loss as the classification objective function. To increase the variability of the training data and to avoid overfitting, we employ a simple data augmentation strategy (Zhou et al., 2020a). Individual graph nodes are randomly translated by a maximum offset. We evaluate the effect of the strength of augmentation and test maximum offsets of 0.1mm, 0.5mm, and 1.0mm. Given the limited amount of training data, data augmentation should be beneficial for improving classification accuracy across different datasets.

All our implementations were done in PyTorch benefiting from the excellent PyTorch Geometric library[7]. We use PyTorch Lightning[8] for ease of implementation of the model and data structures. The code is available on https://github.com/biomedia-mira/medmesh.

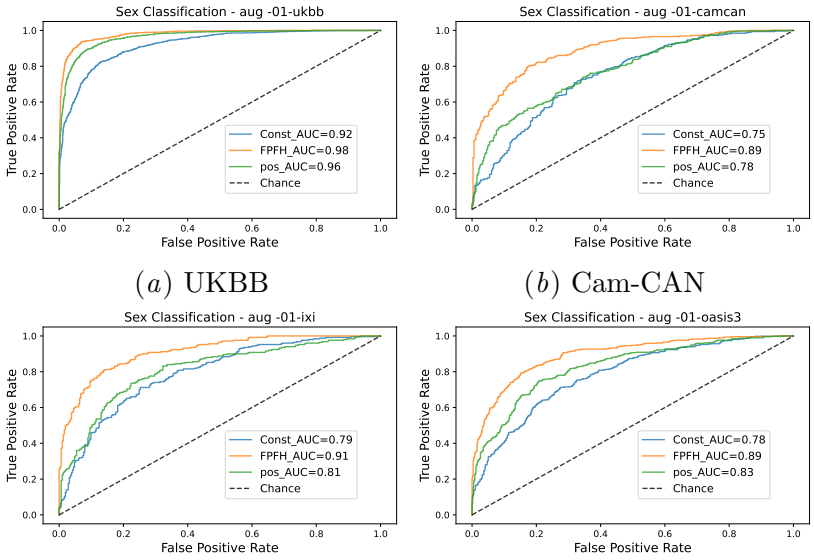
4.2. Task 1: Biological Sex Classification

We use biological sex classification as a proof of concept task which has shown to yield good performance with the advantage that several neuroimaging datasets from different sources are available for extensive testing and evaluation of the effect of different model choices on predictive performance. We use the UKBB data for the model development, with a data split of 70%, 10%, and 20% for training, validation, and testing. The batch size was set to 128, all hidden features set to 32 (both for the convolutional layers and fully connected layers). When using SplineCNN, we set the kernel size to 5 and use the sum aggregation. The maximum number of training epochs was set to 50, and we retain the model with highest validation performance for final evaluation on the test set.

5. https://itk.org
6. https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST
7. https://pytorch-geometric.readthedocs.io/
8. https://www.pytorchlightning.ai/

Effect of node features To evaluate the effectiveness of different node features, in our first set of experiments we employ SplineCNN in the shared convolutional submodel (as these performed well in initial experimentation). We then evaluated classification performance on the UKBB test set, Cam-CAN, IXI, and OASIS-3 using constant, positional, and FPFH node features.

The ROC curves in Figure 2 show that FPFH substantially outperforms other node features on all four datasets. It is worth noting that while positional features perform well on the in-distribution UKBB test set, these features underperform on out-of-distribution test sets. This is due to their reliance on Cartesian coordinates of mesh nodes which do not generalize well due to differences in data acquisition. FPFH, on the other hand, are invariant to the pose of the mesh and show much better generalization across datasets.

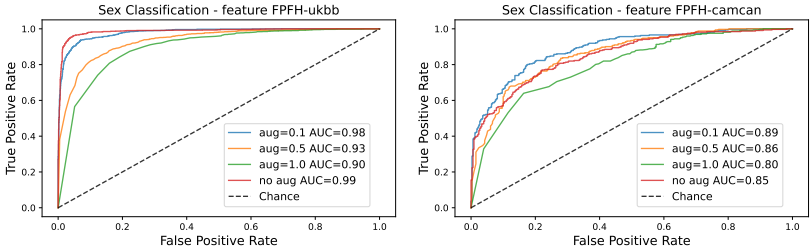


(*a*) UKBB (*b*) Cam-CAN

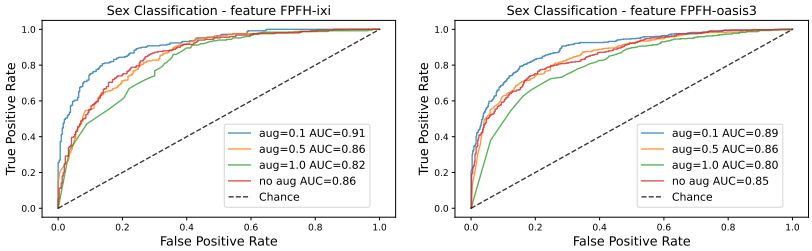
(*c*) IXI (*d*) OASIS-3

Figure 2: ROC curves for sex classification comparing different node features across the datasets UKBB, Cam-CAN, IXI, OASIS-3 using SplineCNN in the submodel.

Effect of data augmentation Next, we evaluate the effect of varying strengths of data augmentation. The maximum offset for the random node translation is varied from 0 (no augmentation), to 0.1, 0.5, and 1.0mm. The ROC curves in Figure 3 demonstrate the benefit of data augmentation on robustness and generalization. The best performance is achieved using data augmentation of 0.1, which increased AUC by 4-5% compared to not using augmentation. While data augmentation slightly decreases the performance on the in-distribution UKBB test set, it substantially improves performance on all out-of-distribution test sets, confirming the importance of adding random perturbations to the training data.



(*a*) UKBB (*b*) Cam-CAN



(*c*) IXI (*d*) OASIS-3

Figure 3: ROC curves showing the effect of data augmentation for sex classification across domains, using SplineCNN as the shared submodel and FPFH as node features.

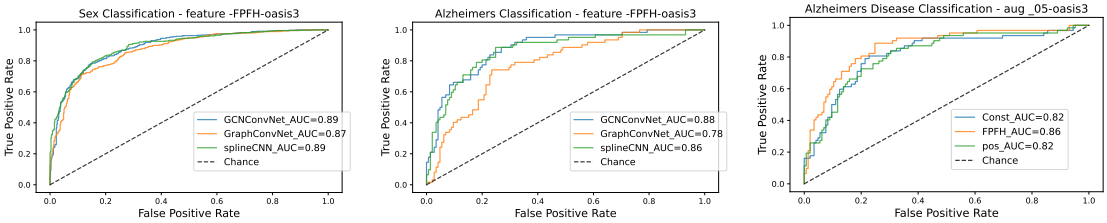
Effect of convolution layer Finally, we evaluate the three different convolutional layers, using FPFH as the node features and data augmentation of 0.1mm. In Figure 4(*a*) we observe similar performance for SplineCNN and GCNConv, closely followed by GraphConv.

4.3. Task 2: Alzheimer's Disease Classification

To confirm whether the above findings hold for a clinically relevant task, we consider Alzheimer's disease (AD) classification on OASIS-3 with a 70%, 10%, and 20% train, validation, and test split. We evaluate the effect of the convolutional layer using a larger amount of data augmentation of 0.5mm due to the smaller amounts of training data. We then also evaluate the effect of node features for AD classification, using SplineCNN in the submodel for consistency with the sex classification experiments. The results are shown in Fig. 4(*b*) and 4(*c*). GCNConv performs slightly better than SplineCNN, with a substantial decrease in performance for GraphConv. FPFH features again outperform other node features.

4.4. Bias Analyses

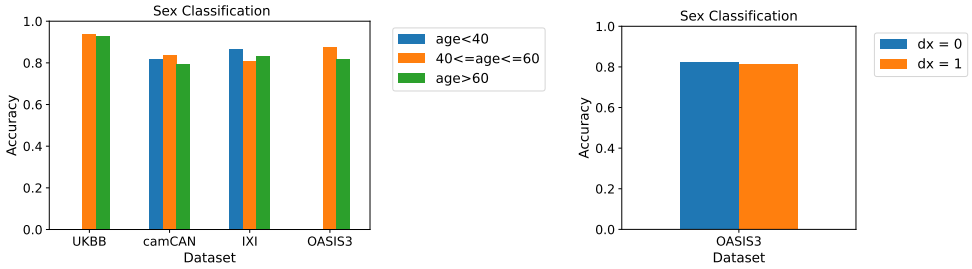
We also investigated potential biases in the predictions in terms of subgroup performance disparities. To this end, we first analyzed the biological sex classification model stratified by age groups. As the training data from UKBB only covers a limited age range between 44 and 73 year old subjects, we wanted to understand whether the performance might degrade for younger subjects. The results shown in Figure 5(*a*), however, suggest that the sex classification model with SplineCNN and FPFH features generalizes well across the



(*a*) Sex - Conv. Layers (*b*) AD - Conv. Layers (*c*) AD - Node Features

Figure 4: Effect of convolution layer for (a) sex and (b) Alzheimer's disease classification. (c) Effect of node features on AD classification. All evaluated on OASIS-3.

entire age range. Both Cam-CAN and IXI contain many subjects in the range of 18 to 40 years. Next, we analyzed whether sex classification may be affected by disease status. Here we looked at the classification performance separately for the group of healthy controls and subjects with Alzheimer's disease. Again, we find no differences in the classification accuracy, suggesting that the sex classification model generalizes well (cf. Figure 5(*b*)).



(*a*) (*b*)

Figure 5: Bias analysis for sex classification using SplineCNN with FPFH features. Classification performance is stratified by (a) age groups and (b) presence of disease.

5. Conclusion

This comparative study evaluated the effect of node features, convolutional layers, and data augmentation on two different tasks and four datasets in medical shape classification with graph neural networks. We find that the use of FPFH features is highly beneficial, substantially improving classification performance on out-of-distribution test data. The FPFH features alleviate the need for data normalization such as mesh alignment due to their pose invariance. We are not aware of earlier studies proposing the use of FPFH in GNNs. We further find that SplineCNN and GCNConv are both viable options for the convolutional layers, yielding comparable performance. We also conclude that data augmentation is

essential in GNNs, in particular, when the amount of training data is limited. We find that stronger data augmentation is beneficial in particular for Alzheimer's disease classification where the training set contained less than 800 samples. AD classification performance was overall promising, and in line with a recent study evaluating data representations (Sarasua et al., 2022). This should be confirmed in future work on other datasets for AD classification such as ADNI.

A limitation of this work is that only relatively simple data augmentation was considered in the form of perturbing the mesh node positions. Recently, there has been work on more advanced data augmentation techniques for graph neural networks (Ding et al., 2022) such as removing a certain number of edges either randomly (Rong et al., 2019) or based on the GNN predictions adjusting the graph in an adaptive manner (Chen et al., 2020) to name a few. In future work, it would be interesting to evaluate the effect of these more advanced techniques for medical shape classification.

Our proposed multi-graph architecture may be useful in other applications which will be investigated in future work.

Acknowledgments

Nairouz Shehata is grateful for the support by the Magdi Yacoub Heart Foundation.

References

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinícius Flores Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Çaglar Gülçehre, H. Francis Song, Andrew J. Ballard, Justin Gilmer, George E. Dahl, Ashish Vaswani, Kelsey R. Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matthew M. Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *CoRR*, abs/1806.01261, 2018. URL http://arxiv.org/abs/1806.01261.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. URL https://arxiv.org/abs/2104.13478.

Nada Chaari, Mohammed Amine Gharsallaoui, Hatice Camgöz Akdağ, and Islem Rekik. Multigraph classification using learnable integration network with application to gender fingerprinting. *Neural Networks*, 151:250–263, 2022.

Deli Chen, Yankai Lin, Wei Li, Peng Li, Jie Zhou, and Xu Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):3438–3445, Apr. 2020. doi: 10.1609/aaai.v34i04.5747. URL https://ojs.aaai.org/index.php/AAAI/article/view/5747.

Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6 (1):1–25, 2019.

Kaize Ding, Zhe Xu, Hanghang Tong, and Huan Liu. Data augmentation for deep graph learning: A survey. *arXiv preprint arXiv:2202.08235*, 2022.

Matthias Fey, Jan Eric Lenssen, Frank Weichert, and Heinrich Müller. Splinecnn: Fast geometric deep learning with continuous b-spline kernels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 869–877, 2018.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

Genxuan Hong, Zhanquan Wang, Taoli Han, and Hengming Ji. Spatiotemporal multi-graph convolutional network for taxi demand prediction. In *2021 11th International Conference on Information Science and Technology (ICIST)*, pages 242–250. IEEE, 2021.

Juan Eugenio Iglesias, Cheng-Yi Liu, Paul M Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, 2011.

Anees Kazi, Shayan Shekarforoush, S Arvind Krishna, Hendrik Burwinkel, Gerome Vivar, Benedict Wiestler, Karsten Kortüm, Seyed-Ahmad Ahmadi, Shadi Albarqouni, and Nassir Navab. Graph convolution based attention model for personalized disease prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 122–130. Springer, 2019.

Byung-Hoon Kim, Jong Chul Ye, and Jae-Jin Kim. Learning dynamic graph representation of brain connectome with spatio-temporal attention. *Advances in Neural Information Processing Systems*, 34:4314–4327, 2021.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, 2019.

Chuan Li. Openai's gpt-3 language model: A technical overview, Sep 2020. URL https://lambdalabs.com/blog/demystifying-gpt-3/.

Junying Li, Deng Cai, and Xiaofei He. Learning graph-level representation for drug discovery. *arXiv preprint arXiv:1709.03741*, 2017.

Karla L Miller, Fidel Alfaro-Almagro, Neal K Bangerter, David L Thomas, Essa Yacoub, Junqian Xu, Andreas J Bartsch, Saad Jbabdi, Stamatios N Sotiropoulos, Jesper L R

Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Thomas W Okell, Peter Weale, Iulius Dragonu, Steve Garratt, Sarah Hudson, Rory Collins, Mark Jenkinson, Paul M Matthews, and Stephen M Smith. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, 19(11):1523–1536, 2016. ISSN 1097-6256. doi: 10.1038/nn.4393.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4602–4609, 2019.

John C Morris. The clinical dementia rating (cdr): Current version and. *Young*, 41:1588–1592, 1991.

Brian Patenaude, Stephen M Smith, David N Kennedy, and Mark Jenkinson. A bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3): 907–922, 2011.

Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang. Dropedge: Towards deep graph convolutional networks on node classification. *arXiv preprint arXiv:1907.10903*, 2019.

Radu Bogdan Rusu, Zoltan Csaba Marton, Nico Blodow, and Michael Beetz. Learning informative point classes for the acquisition of object model maps. In *2008 10th International Conference on Control, Automation, Robotics and Vision*, pages 643–650. IEEE, 2008.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009.

Ignacio Sarasua, Sebastian Pölsterl, and Christian Wachinger. Hippocampal representations for deep learning on alzheimer's disease. *Scientific reports*, 12(1):1–13, 2022.

Meredith A. Shafto, Lorraine K. Tyler, Marie Dixon, Jason R. Taylor, James B. Rowe, Rhodri Cusack, Andrew J. Calder, William D. Marslen-Wilson, John Duncan, Tim Dalgleish, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurology*, 14(1):204, 2014.

Jonathan Shlomi, Peter Battaglia, and Jean-Roch Vlimant. Graph neural networks in particle physics. *Machine Learning: Science and Technology*, 2(2):021001, 2020.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

Jason R. Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A. Shafto, Marie Dixon, Lorraine K. Tyler, Richard N. Henson, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 144:262–269, 2017. doi: https://doi.org/10.1016/j.neuroimage.2015.09.018.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, 2010.

Wei Wang, Junyang Chen, Yushu Zhang, Zhiguo Gong, Neeraj Kumar, and Wei Wei. A multi-graph convolutional network framework for tourist flow prediction. *ACM Transactions on Internet Technology (TOIT)*, 21(4):1–13, 2021.

Kyle Wiggers. Openai's massive gpt-3 model is impressive, but size isn't everything, Jun 2020. URL https://venturebeat.com/ai/ai-machine-learning-openai-gpt-3-size-isnt-everything/.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Jiajun Zhou, Jie Shen, and Qi Xuan. Data augmentation for graph classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2341–2344, 2020a.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020b.