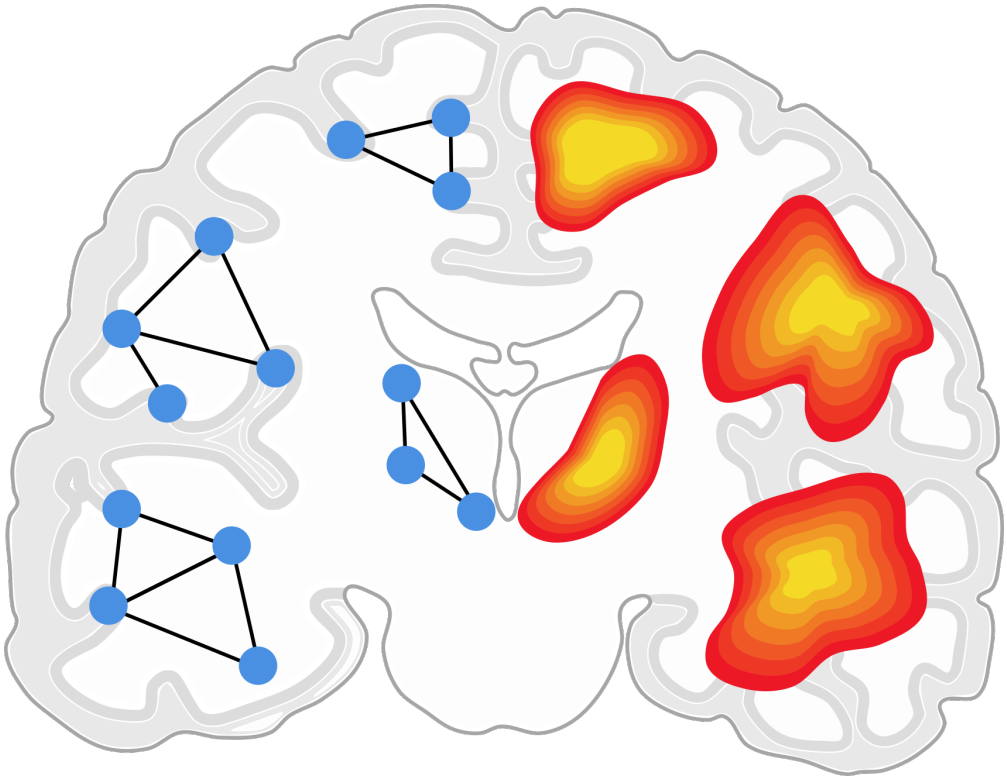
CHALMERS
UNIVERSITY OF TECHNOLOGY



Saliency mapping of RS-fMRI data in GCNs for sex and brain age prediction

Identifying important functional brain networks using explainability in Graph Convolutional Networks

Master's thesis in Physics

KEVIN ANDERSSON & ERIC LINDGREN

DEPARTMENT OF PHYSICS

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

Master's thesis 2021

Saliency mapping of RS-fMRI data in GCNs for sex and brain age prediction

Identifying important functional brain networks using explainability in Graph Convolutional Networks

Kevin Andersson & Eric Lindgren

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Physics
Chalmers University of Technology
Gothenburg, Sweden 2021

Saliency mapping of RS-fMRI data in GCNs for sex and brain age prediction
Identifying important functional brain networks using explainability in Graph Convolutional Networks
Kevin Andersson & Eric Lindgren

Supervisor: Alice Deimante Neimantaite & Lisa Sjöblom, Syntronic Research and Development AB
Examiner: Giovanni Volpe, Department of Physics, University of Gothenburg

Master's Thesis 2021
Department of Physics
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone: +46 31 772 1000

Cover: Rendition of a human brain with fictional functional brain networks interpreted as graphs.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2021

Saliency mapping of RS-fMRI data in GCNs for sex and brain age prediction
Identifying important functional brain networks using explainability in Graph Convolutional Networks
Kevin Andersson & Eric Lindgren
Department of Physics
Chalmers University of Technology

Abstract

Insight into how biological sex and healthy ageing affects the human brain are important for an increased understanding of the brain. Healthy ageing insights are also useful for clinical applications, for instance in identifying unhealthy ageing due to neurodegenerative disease. To this end, several studies in the last few years have used machine learning methods on neuroscientific data to predict subject sex and brain age. One particularly interesting approach has been to represent functionally connected networks in the brain as graphs, and apply Graph Convolutional Networks (GCNs). To investigate which functional brain networks are connected with sex and age, we develop and analyse GCN-based models that predict sex and age from resting-state fMRI data. The analysis of the models is done using saliency mapping techniques that give insight into which functional brain networks in the data are relevant for the predictions. With this approach, we obtain a sex prediction accuracy of up to 79 % and an age prediction MAE of 5.9 years. Furthermore, we find indications that the Somatomotor Medial Network and the cerebellum are among the more important functional brain networks for predicting sex and brain age.

Keywords: machine learning, supervised learning, GNN, GCN, explainability in AI, graph theory, population graphs, brain age, sex, functional connectivity, resting-state fMRI, saliency mapping.

Acknowledgements

This Master's thesis project is a collaboration between Syntronic AB, University of Gothenburg and Karolinska Institutet, and hence there are a number of people for which some thanks are in order. First of all, we would like to thank our supervisors, Alice Deimante Neimantaite and Lisa Sjöblom at Syntronic Research and Development AB. Alice and Lisa, your expertise, both in the domains of machine learning and in neuroscience, your support, and all the interesting discussions we have had, have been the cornerstone of this project. Next, our thanks go out to Joana Pereira, assistant professor, and Mite Mijalkov, post-doctoral researcher, at Karolinska Institutet (KI) for providing us with the data used in this thesis, as well as helping us interpret the results from a neuroscientific perspective. We would also like to thank Giovanni Volpe, professor at the Physics Department at University of Gothenburg, for providing guidance to the work and for being the examiner for this project. Finally, we would like to thank Syntronic AB for providing the opportunity for this fantastic collaboration, and for making the Master's thesis project possible in the first place.

Kevin Andersson & Eric Lindgren, Gothenburg, June 2021

Declaration of contributions

The data used in this thesis has been procured from the UK Biobank project, `https://www.ukbiobank.ac.uk/`, an extensive biomedical database for public health research containing data from over half a million participating subjects. Examples of types of data that are included in the biobank are subject sex, age, cognitive abilities, disease history, but also more specific information such as MRI and fMRI brain scans. Mite Mijalkov, post-doctoral researcher at KI, has provided us with the 21-node fMRI brain graphs for the approximately 35000 subjects used in this thesis, after which we have performed minimal preprocessing and stratification. Joana Pereira, assistant professor at KI, has contributed with the information in Table 3.1, in which each of the 21 nodes are mapped to their corresponding functional brain network. Finally, Figure 3.1 has been adapted from a figure from UK Biobank Brain Imaging Online Resources, `https://www.fmrib.ox.ac.uk/ukbiobank/`, specifically at `https://www.fmrib.ox.ac.uk/ukbiobank/netjs_d25/`.

Table of Contents

List of Figures

List of Tables

Acronyms

ANN Artifical Neural Network. 2, 10

BCE Binary Cross Entropy. xi, 23, 24, 26, 27

BG Basal Ganglia. 14

CB1 Cerebellum. 14, 29–31, 34, 35

CB2 Cerebellum. 14, 27–31, 34–36

CNN Convolutional Neural Network. 7, 9, 38

DAL Dorsal Attention Left. 14

DAR Dorsal Attention Right. 14

DMN Default Mode Network. 14, 29–31, 34–36

DTI Diffusion Tensor Imaging. 1

EEG Electroencephalography. 1

fMRI functional Magnetic Resonance Imaging. ix, 1–3, 11, 13, 14, 17, 33, 37

FP Fronto-parietal. 14

GAT Graph Attention Network. 3

GCN Graph Convolutional Network. 2, 3, 7, 9, 10, 15, 16, 33, 37, 38

GIN Graph Isomorphism Network. 3, 11

GNN Graph Neural Network. 2, 11

MAE Mean Absolute Error. xi, 23, 25, 33

MCC Matthews Correlation Coefficient. xi, 23, 24

MEG Magnetoencephalography. 1

MRI Magnetic Resonance Imaging. 1, 3, 10, 13, 33, 37

MSE Mean Squared Error. xi, 23, 25, 26, 29, 33

PET Positron Emission Tomography. 1

PL Prefrontal Lateral. 14, 27–29, 31, 35

PMC Posteromedial Cortex. 14, 27–29, 31, 35

PMN Primary Motor Network. 14

PSS Primary Somatosensory. 14

ReLU Rectified Linear Unit. 16

RS-fMRI Resting-State Functional Magnetic Resonance Imaging. 1, 3, 11, 37

SMM Somatomotor Medial. 14, 27, 29–31, 34–37

SMN Sensorimotor Network. 35

SN Salience Network. 14, 29–31, 34, 35

SSN Somatosensory Network. 14, 35

SVM Support Vector Machine. 3, 35

TM Temporal Medial. 14

TS Temporal Superior. 14, 28, 29

VA Ventral Attention. 14

VL Visual Lateral Network. 14

VM Visual Medial. 14

VV1L Visual V1 Lateral. 14

VV1M Visual V1 Medial. 14, 28, 34, 35

1. Introduction

In recent years, the interface between neuroscience and machine learning has been an active area of research. By combining large amounts of neuroscientific data with recent advances in machine learning, interesting insights into the human brain have been obtained [1]–[3]. Another fruitful approach has been to apply the framework of graph theory in neuroscience, where the human brain is represented as a graph, and analysed using graph theoretic techniques [4], [5]. In this project, we have combined these approaches by applying graph-based machine learning models on neuroscientific data in order to understand healthy ageing, and differences between sexes, in the human brain. The motivation for this study is twofold; generally, to understand how the human brain potentially differs with age and sex, but also specifically since understanding the healthy ageing of the human brain may aid in identifying unhealthy ageing, perhaps due to underlying neurodegenerative disease [6].

1.1 Neuroscience of human brain connectivity

The human brain can be analysed by studying how the connections between various parts of the brain differ with sex and age. These connections can either be structural or functional. Structural connections refer to different parts of the brain being anatomically connected, whilst functional connections are slightly more abstract in that two brain regions are considered connected if their activation is correlated in time [7]. The structural connections of the brain can be obtained through methods such as MRI and DTI, whereas functional connections are studied through functional-MRI (fMRI), PET, EEG and MEG [8]. Mapping these connections throughout the brain yield structural and functional brain networks, respectively. The functional networks can further be grouped together to form specialised networks. The largest of these is the default-mode network, which is primarily active when the brain experiences no external stimuli and for example considered to be responsible for mind wandering [9].

Functional brain networks can be derived from either task-specific or resting-state fMRI (RS-fMRI), depending on the state of the subject when the measurement is carried out. Task-specific networks are mapped whilst the subject is engaged in some activity, for instance memory tasks [10]. In contrast, resting-state functional networks are mapped when the subject is not exposed to any external stimuli, i.e., is in a resting state. In Biswal et. al. [11], it was observed that such resting-state functional networks are affected by sex and age, and that these effects where observed to be consistent across a large group of subjects.

1.2 Machine learning and graph theory in neuroscience

One possible framework for analysing functional brain networks is graph theory. The networks are interpreted as graphs, and properties of these graphs can then be extracted using graph measures. This approach has previously been applied on functional brain networks with promising results. In Chan et al. [4], graph measures where used to analyse the functional connections between various specialised brain networks, with the graph theoretical measure of segregation among the specialised networks being correlated to ageing. However, there still exists a need for more complex analysis to further understand the functioning of the human brain with regards to age and sex. For instance, it has been observed that patients suffering from neurodegenerative diseases such as dementia exhibit gaps in their estimated brain age compared to their actual age [6]. Thus, if one could develop efficient methodologies for estimating brain age, such methods could be used to identify and analyse unhealthy ageing due to underlying neurodegenerative disease.

Inspired by recent advances in machine learning, Artifical Neural Networks (ANN) have previously been utilised to obtain accurate and complex models for predicting sex and brain age [2], [12], [13]. ANNs typically excel at finding complex patterns in large data sets. However, classical ANNs are heavily dependent on the input data forming ordered tensors. This is not the case for graph structured data, since graphs by definition are node order invariant. Thus, if one wants to combine ANNs with the promising approach of analysing functional brain networks as graphs, one could turn to a class of ANNs known as Graph Neural Networks (GNN) which operate directly on the graph structure. Graph Convolutional Networks (GCN) are a subclass of GNNs which generalise the concept of convolutions to the graph domain. GCNs have shown promising results when applied to graph data in general [14]–[16], and specifically for the task of predicting neurodegenerative disease in human brains based on graphs derived from structural-MRI data [17]. Motivated by these results, GCNs applied on large fMRI data sets could be used to develop accurate models for predicting sex and brain age, which in turn could be analysed to gain further insight into the functioning of the human brain.

1.3 Aim of the project

The aim of this thesis is to investigate which resting-state functional brain networks are important for age and sex prediction. To this end, machine learning models based on GCNs are developed and analysed using saliency mapping techniques. A saliency map determines the importance of, for instance, individual pixels in an image or nodes in a graph, and will in this thesis reveal which functional brain networks are important for the predictions.

1.3.1 Demarcations

Since both neuroscience and machine learning are two large and active fields of research, some demarcations must be set for this thesis. Regarding the data to be analysed, one such demarcation is that only RS-fMRI data will be investigated, which is a delimitation both in the sense that task-specific fMRI will not be included, but also that other data modalities such as MRI will not be utilised. Another demarcation is that in the data used in this thesis, the connections defined by correlations from the fMRI scans are aggregated into graphs with 21 connected nodes, where each node represents a smaller functional network of brain regions. This is a delimitation since using data with other dimensionalities could perhaps be beneificial.

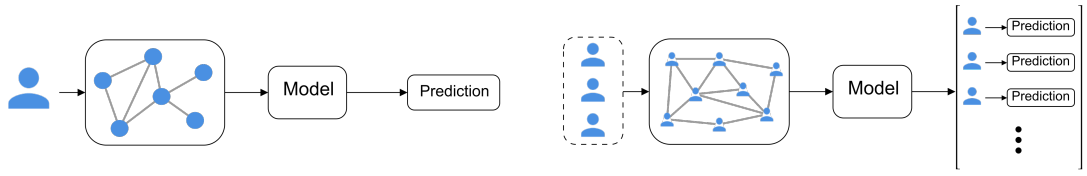
A demarcation regarding the machine learning is that only models based on GCNs, with the exception of a baseline model for comparison, will be considered. There exist other types of models and architectures that could be investigated for the aim of the thesis. For instance other GNNs, including Graph Attention Networks (GAT) [18] and Graph Isomorphism Networks (GIN) [19]. More generally, one could also study the usage of other kinds of machine learning models such as Support Vector Machines (SVM) [1].

1.4 Structure of the report

The first chapter will include an overview of general graph theory and Graph Convolutional Networks, as well as the theoretical foundation for a saliency mapping technique. In the second chapter, all models used in the thesis will be presented, together with other necessary methods that have been developed, for instance to preprocess the data and be able to train the networks. Thereafter, the results will be presented, including model performance and saliency maps. Lastly, the thesis will be concluded with a discussion, conclusion and outlook.

2. Theory

The approach of using graph theory to model networks is abundant in many areas of sciences, ranging from engineering to biology [20]. From the perspective of machine learning, there exist many graph related tasks, with two common ones being *graph classification* and *node classification* [21], [22]. In graph classification, the task is to classify a graph as belonging to one of several classes, whilst in node classification the task is to classify each node in the graph individually. These two problems can be generalised to graph and node regression, with the aim of predicting a continuous variable instead of a class. Graph and node classification/regression can be referred to as graph and node prediction.



(a) Graph prediction **(b)** Node prediction

Figure 2.1: Schematic representations of graph and node prediction.

In the context of neuroscience, graph and node prediction may correspond to predicting some property of a subject, for instance sex or age. Graph prediction is relevant when each subject is associated with a graph, which in this thesis will correspond to functional networks in a subject's brain. A schematic visualisation of graph prediction can be seen in Figure 2.1a. Node prediction is relevant in, for instance, *population graphs*, where each node represents a single subject. See Figure 2.1b for a representation of node prediction in population graphs.

2.1 General graph theory

Let $\mathcal{G}$ be a graph, associated with a set of nodes V and a set of edges E. The edges form the connections between nodes, and for a pair of nodes (i, j) the corresponding edge may be denoted e_{ij}. The neighbours $N(v)$ to a certain node $v \in V$ are defined as the nodes $u \in V$ that are connected to v, i.e., $e_{vu} \in E$. Each node v may also be associated with *node features* x_v. x_v may represent some state or information about node v, which in practice can take the form of a vector of numbers [22].

A graph $\mathcal{G}$ is considered to be *undirected* if all edges in the graph are undirected, $e_{ij} \equiv e_{ji}$, and *directed* if all edges are directed, $e_{ij} \not\equiv e_{ji}$. The edges in $\mathcal{G}$ may be

succinctly summarised in an *adjacency matrix* A, defined as

$$A_{ij} = \begin{cases} 1, & \text{if } e_{ij} \in E \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

From definition (2.1) it follows that undirected graphs have a symmetric adjacency matrix, $A_{ij} = A_{ji}$. From the adjacency matrix, the degree matrix D is defined as

$$D_{ij} = \begin{cases} \sum_k A_{ik}, & \text{if } i = j \\ 0, & \text{otherwise,} \end{cases} \tag{2.2}$$

which describes each node's *degree*, i.e., the number of edges connected to each node.

Moreover, each edge in $\mathcal{G}$ may be associated with a weight w_{ij}, and $\mathcal{G}$ is then denoted as a *weighted* graph. The adjacency matrix becomes $A_{ij} = w_{ij}$, and the degree matrix now describes the *strength* of each node, which is defined as the sum of the weights of all edges connected to the node. Weights in a weighted graph may be interpreted differently depending on the application [23]. They can be seen as the connection strength between devices in a cellular network, the cost of a ticket along a route in a train network, the flow of goods in a logistics setting etc. All graphs referenced in the remainder of this thesis will be weighted and undirected unless otherwise specified.

2.2 Population graphs and multiplex graphs

In a population graph, each node corresponds to an individual, and the edges relate all individuals to each other. One way of calculating the weights in a population graph is by using a similarity measure. The similarity measure greatly shapes the resulting population graph, and should be chosen with the specific application in mind. One common approach is to define a similarity measure σ as a distance measure D inverted by a kernel K. The similarity measure between two data points x_1 and x_2 may then be written as

$$\sigma(x_1, x_2) = K\left(D\left(x_1, x_2\right)\right). \tag{2.3}$$

Another type of graph, relevant for this thesis, is a multiplex graph. A multiplex graph is a layered graph where each layer can represent a different type of interaction between nodes [24]. One example of a multiplex graph with the same nodes in all layers is a graph with cities as nodes and different means of transportation as edges. Edges representing highways and railways make up two different graphs, but they might be combined into a multiplex graph since they share the same nodes, i.e., the cities. See Figure 2.2 for a visual representation of a two-layered multiplex graph, in which the nodes are fixed, but the edges are different in each layer.

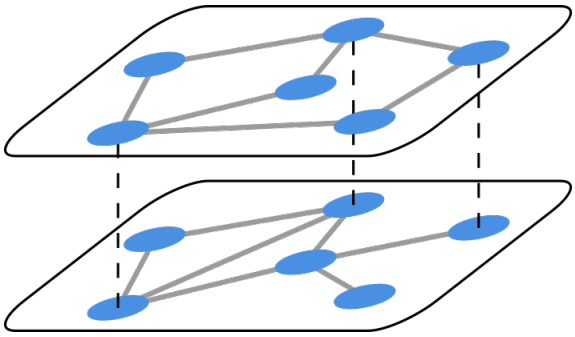


Figure 2.2: An example of a two-layered multiplex graph, with the same nodes in each layer but different edges.

2.3 Graph Convolutional Networks

Graph Convolutional Networks (GCN) are a class of neural networks that generalise the notion of convolutions from grid data to graph structured data [16]. Regular convolutional operations operate on structured grids of data, for instance the pixels in an image, where each grid point is only connected to its adjacent neighbours. This may not be the case for graph structured data, where any pair of nodes can be connected. Generalising the convolutional operator to the graph domain enables a GCN to utilise information about the graph structure, in the same way a CNN can utilise information about structures in e.g. an image [16].

The theoretical foundations for GCNs will be explained in three parts. First, the concept of convolutions in the graph domain, and how they are implemented in a computationally efficient manner. Then, a layer-wise propagation rule for a graph convolutional layer based on the implementation of convolutions, and finally a heuristic interpretation known as *message passing*.

2.3.1 Efficient convolutions in the graph domain

Let $\mathcal{G}$ be an undirected graph, with N nodes and adjacency matrix A. The graph Laplacian L is defined as $L = D - A$, and may be normalized according to

$$L_{\text{norm}} = D^{-1/2}LD^{-1/2} = I_N - D^{-1/2}AD^{-1/2}, \tag{2.4}$$

where I_N is the $N \times N$ identity matrix. With $L = L_{\text{norm}}$, the convolution of a signal $x \in R^N$ defined on the nodes of $\mathcal{G}$ with a filter $g_\theta = \text{diag}(\theta)$ parametrized by $\theta \in R^N$, can in the Fourier domain be written as

$$g_\theta * x = U g_\theta U^T x, \tag{2.5}$$

where U is an orthogonal matrix containing the eigenvectors of L [14]. Performing convolutions using Equation (2.5) may, however, be computationally intractable in practice, partly because multiplication with U is $\mathcal{O}(N^2)$, and partly because calculating U requires the eigendecomposition of L, which may be very computationally

expensive for large graphs [14]. Kipf et al. [14] proposes a solution to this problem, in which the convolution is approximated with an expansion in Chebyshev polynomials $T_k(x)$, as,

$$g_{\theta'} * x \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{L})x, \tag{2.6}$$

where $\tilde{L} = \frac{2}{\lambda_{\max}} L - I_N$, and $\lambda_{\max}$ is the largest eigenvalue of L. θ'_k is here a vector of Chebyshev coefficients, which in the context of machine learning corresponds to trainable parameters. The interpretation of Equation (2.6) is that the convolution is K^{th} order local, i.e. that a convolution considers each node and its K^{th} order neighbors, the neighbours that are at most K steps away in the graph.

2.3.2 Layer-wise propagation rule

By setting $K = 1$ and a single trainable parameter $\theta = \theta'_0 = -\theta'_1$, the convolution in Equation (2.6) can be further simplified as

$$g_{\theta'} * x \approx \theta \left(I_N + D^{-1/2} A D^{-1/2} \right) x, \tag{2.7}$$

[14]. Setting $K = 1$ limits the convolution to only consider a node ($k = 0$) and its closest neighbours ($k = 1$). Implementing convolutions as in Equation (2.7) may, however, lead to exploding or vanishing gradients, since the eigenvalues of $I_N + D^{-1/2} A D^{-1/2}$ lies in the range $[0, 2]$. This problem can be avoided using a *renormalization trick*:

$$I_N + D^{-1/2} A D^{-1/2} \rightarrow \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2}, \tag{2.8}$$

where $\tilde{A} = A + I_N$ and $\tilde{D} = \sum_j \tilde{A}_{ij}$.

The convolutional operation may now be generalized to a signal $X \in \mathbb{R}^{N \times C}$, defined on the N nodes of the graph and with a C-dimensional feature vector for each node. The signal X thus has C input channels. Furthermore, let the convolution apply F filters to the C input channels. The application of such a convolution can then be written

$$Z = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} X \Theta, \tag{2.9}$$

with $\Theta \in \mathbb{R}^{C \times F}$ being a matrix of filter parameters and $Z \in \mathbb{R}^{N \times F}$ being the convolved signal. This propagation rule in Equation (2.9), when paired with an activation function σ such that $H = \sigma(Z)$, defines a graph convolutional layer.

Recall that, in order to arrive at Equation (2.9), the Chebyshev polynomial expansion in Equation (2.6) was truncated at $K = 1$. By truncating the Chebyshev polynomial, the convolution operator only considers each node and its closest neighbours, and becomes a linear function in L. These two points may seem to be limitations in the representational strength of the graph convolutional layer, however, this is not necessarily the case [14]. Firstly, larger neighbourhoods can be convolved over by stacking multiple layers, with the first layer considering a node and its neghbours, the second layer considering the neighbours' neighbours and so on. Secondly, keeping the convolutional operator linear in L actually makes it more flexible, since it is

not dependant on the explicit parametrization given by the Chebyshev polynomials [14]. Combined, a deep GCN consisting of several stacked graph convolutional layers, paired with possibly non-linear activation functions σ, can still model a rich class of convolutional filter functions, whilst keeping the computational costs low.

2.3.3 Message passing interpretation of GCNs

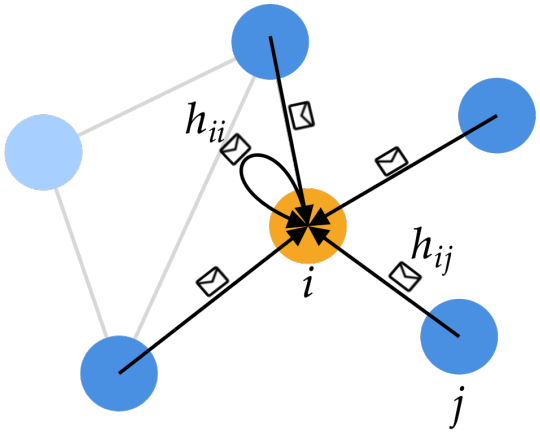


Figure 2.3: A schematic representation of the message passing interpretation of a graph convolutional layer. The messages h_{ij} are sent from node j to node i.

Heuristically, a graph convolutional layer can be seen to perform a sort of *message passing*, in that each node receives messages (aggregates features) from all of its closest neighbours. This can be seen by studying the propagation rule in Equation (2.9) for a single node i, a single feature channel $C = 1$ and a single filter channel $F = 1$, and explicitly rewriting it as a sum:

$$\begin{aligned} z_i &= \theta\left(\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}\right)_{ii} x_i + \sum_{j\neq i}\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}\right)_{ij} x_j\right) \\ &= \theta\left(h_{ii} + \sum_{j\neq i} h_{ij}\right). \end{aligned} \tag{2.10}$$

z_i and x_i corresponds to the activation and feature for node i, respectively. h_{ij} can be interpreted as a normalised message, communicating the state x_j weighted by the normalised connection between node j and i. The activation z_i can then be seen as an aggregation of normalised messages from all neighbouring nodes, h_{ij}, and itself, h_{ii}. See Figure 2.3 for a visual representation of how the messages are passed to node i from it's neighbours. In this manner information can flow through the graph, and by stacking multiple graph convolutional layers the information can spread over successively larger neighbourhoods. With the message passing interpretation, a deep GCN can be seen to draw upon the informational flow patterns in the graph. This enables it to learn from the graph structure, similarly to how a deep CNN can extract information from structures in an image.

2.4 Zorro – an algorithm for saliency mappings in GCNs

Zorro is an algorithm for determining which nodes and features are important for a trained GCN model performing node classification on a graph with adjacency matrix A and feature matrix X [25]. The model is denoted $\Phi_n(X, A)$ and takes A and X as input, and gives a prediction for a node n in the graph. The general idea is to replace the feature matrix X by noise and then reintroduce nodes and features to successively make the model prediction similar to the original. The reintroduced nodes V and features F are referred to as an explanation $\mathcal{S} = \{V, F\}$. The noisy signal $Y_\mathcal{S}$ is obtained by applying noise to X through element-wise multiplication with a masking matrix S, representing $\mathcal{S}$, and a matrix of random noise Z, according to

$$Y_\mathcal{S} = X \odot S + Z \odot (1 - S), \quad Z \sim \mathcal{N}. \tag{2.11}$$

1 is a matrix of ones, $\odot$ represents element-wise multiplication and the noise distribution $\mathcal{N}$ is chosen to match the distribution of features over the nodes, i.e, the distribution of X. The prediction of the model Φ on the masked data is then $\Phi_n(Y_\mathcal{S}, A)$. To evaluate an explanation $\mathcal{S}$, i.e., if the unmasked nodes and features are important for the prediction of a specific node, the fidelity of the explanation is calculated as

$$\mathcal{F}(\mathcal{S}) = \mathbb{E}_{Y_\mathcal{S}|Z\sim\mathcal{N}}\left[\mathbb{1}_{\Phi_n(X,A)=\Phi_n(Y_\mathcal{S},A)}\right]. \tag{2.12}$$

The fidelity of an explanation is a measure of how likely it is for a model with a noisy signal $Y_\mathcal{S}$ to give the same prediction as a model with the original signal X.

For each node n, the algorithm begins with an empty explanation $S_n = \{\emptyset, \emptyset\}$, and thus a completely random input signal according to Equation (2.12). Then, the node or feature that increases the fidelity of the explanation the most is added to the explanation. Nodes and features are iteratively added in this manner until the fidelity of the explanation exceeds a hyperparameter τ. The resulting explanation indicates which nodes and features were important to make the prediction for node n, yielding a saliency map detailing the importance of each node. The algorithm can then be repeated for all the nodes in the graph.

2.5 Similar works: the intersection of GCNs and neuroscience

There have been numerous papers published in the last few years applying graph neural networks in neuroscience, with similar aims as that of this thesis; to develop accurate classifiers for brain age and sex prediction, and analyse them using saliency mapping techniques. For age prediction, notable works include Stankevičiūtė et al. [12] and Amoroso et al. [2], which both utilise structural MRI data. Stankevičiūtė et al. apply GCNs to a population graph of subjects, and Amoroso et al. use an ANN with graph measures as input features to predict age. The problem of sex prediction is approached in Arslan et al. [13], who use GCNs, and in the work by

Kim and Ye [26], who use a variant of GNN known as GINs. Both of these studies apply their models to RS-fMRI data. Furthermore, [13] and [26] also analyse the gradients within the models, with the goal of identifying which functional brain networks are related to biological sex.

There are three main differences that sets this thesis apart from the work in [2], [12], [13], [26]. Firstly, resting-state fMRI data is used for age prediction. Secondly, relevant functional brain networks for age prediction are investigated. Finally, saliency maps of the relevant functional brain networks are obtained using methods that only considers the input-output behaviour of the models.

3. Method

In this chapter, we present the methodology for how accurate models for predicting age and sex were developed, and how they were analysed using saliency mapping techniques in order to gain insight into which functional brain networks in the data are related to age and sex. The chapter can be summarised as follows: firstly, we present how the fMRI data was preprocessed into graphs suitable for graph neural networks. Secondly, the models developed for graph and node prediction are presented, followed by the methods used for analysing and creating saliency maps from the models. See Appendix A for more details on the exact model architectures used.

3.1 From RS-fMRI scans to graphs

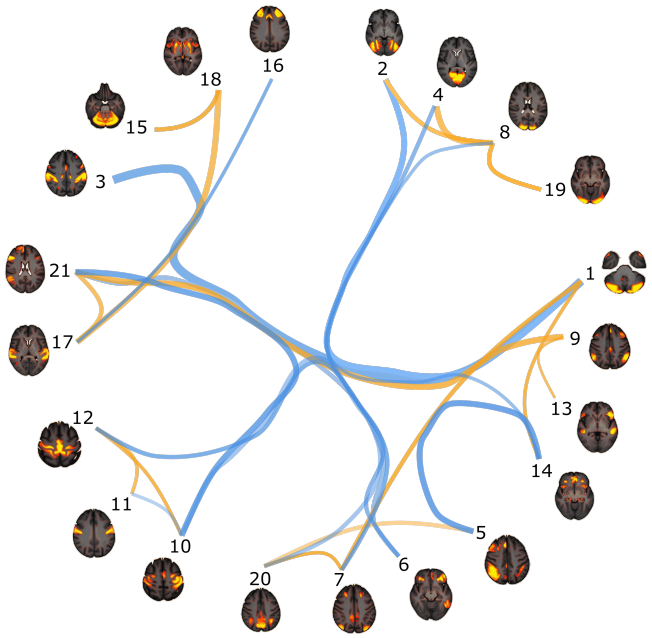


Figure 3.1: A schematic visualisation of a network of functional brain networks. Blue and orange edges represent positive and negative correlation between brain networks, respectively. The individual functional brain networks are represented as human brains with fMRI activations (orange), along the circumference of the circle. See Table 3.1 for the names of the functional brain networks. Image adapted from example at UK Biobank Brain Imaging Online Resources [27].

The graphs of functional brain networks used in this thesis are derived from fMRI. In short, functional MRI (fMRI) is a technique for measuring the neuronal activity in the human brain [7]. In a measurement, the activations of various parts of the subject's brain are measured as time series. The correlations between these time

series can then be calculated and formed into a network. Different sub-networks in the obtained fMRI brain network can often be associated with different functionalities of the brain. One example of such a functional brain network is the Default Mode Network (DMN), which handles memory processing and mind wandering [9]. The regions that make up the sub-networks do not necessarily have to be physically close together nor directly anatomically connected.

Table 3.1: A list of which node corresponds to which functional brain network in the data from the UK Biobank. See Figure 3.1 for a representation of each functional brain network. Each node will in the remainder of this thesis be referred to by the abbreviation for its functional brain network.

Number	Network	Abbreviation
1	Cerebellum	CB1
2	Visual Lateral Network	VL
3	Somatosensory Network	SSN
4	Visual Medial	VM
5	Dorsal Attention Right	DAR
6	Dorsal Attention Left	DAL
7	Fronto-parietal	FP
8	Visual V1 Medial	VV1M
9	Default Mode Network	DMN
10	Primary Somatosensory	PSS
11	Primary Motor Network	PMN
12	Somatomotor Medial	SMM
13	Ventral Attention	VA
14	Salience Network	SN
15	Cerebellum	CB2
16	Prefrontal Lateral	PL
17	Temporal Superior	TS
18	Basal Ganglia	BG
19	Visual V1 Lateral	VV1L
20	Posteromedial Cortex	PMC
21	Temporal Medial	TM

This thesis utilises fMRI data obtained form the UK Biobank for roughly 35000 subjects, of both sexes, varying between 45 and 80 years of age [28]. The fMRI data had been preprocessed and consisted of correlations between 21 different functional brain networks. The correlation between these 21 networks thus constituted a network of networks, and is interpreted as a graph where each node corresponds to one functional network and each edge the correlation between networks. A visualisation of such a brain graph is given in Figure 3.1, in which the various nodes and the functional brain networks they represent form the outer circle, with the coloured lines connecting the nodes being the edges in the graph. The lines are coloured according to the edge weights, with blue and orange lines having positive and negative weights, respectively. For a list of what functional brain network each node corresponds to, and its abbreviation, see Table 3.1.

3.1.1 Treating negative edge weights

A brain graph with both positive and negative weights may be problematic. Specifically, when normalising the adjacency matrices according to Equation (2.8), the risk of dividing by zero becomes imminent. There are several ways to handle this problem. One alternative is to only study all positive or all negative connections; another might be to take the absolute value of all connections. We, however, decided to split the negative and positive connections into two separate graphs and thus form a multiplex graph. From a practical point of view this was implemented by creating a block diagonal adjacency matrix with all positive connections in the upper block and all negative connections in the lower block, see Figure 3.2. The connections in the positive block that were negative in the original adjacency matrix A were replaced with zeros, and vice versa in the negative block.

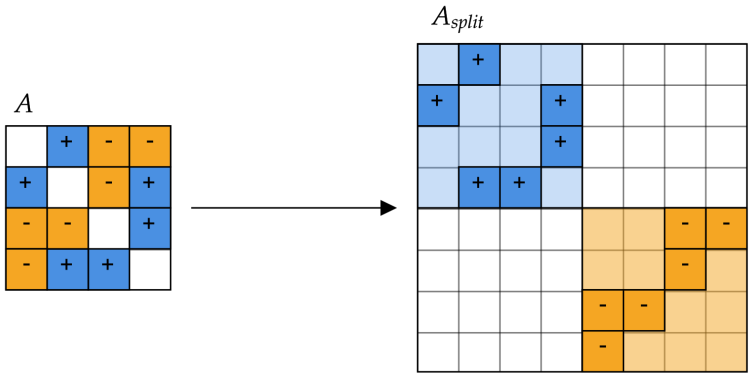


Figure 3.2: An example of how negative values in an adjacency matrix A were handled. The negative values (orange) were extracted into the lower diagonal block of A_{split}, and replaced with zeros in the upper, positive block (blue).

3.2 Models for graph prediction

Having described the preprocessing of the brain graphs, we now turn to the models used to perform graph prediction for individual subjects. Two models will be presented, a baseline regression model and a GCN model, for the tasks of predicting subject age and sex.

3.2.1 Baseline

To validate how well the different GCN-models performed, a baseline model had to be introduced for comparison. In this study, the baseline model consisted of a regression model where the connections between all of the nodes were regarded as separate input features. The adjacency matrix for an individual subject could then be viewed as a high dimensional data point, and the model thus aimed to fit a hyper plane to either separate the classes or predict a continuous variable. For a schematic view of the model, see Figure 3.3. Note that the output layer has either two softmax-activated neurons in the case of sex classification, or one neuron without activation in the case of age regression.

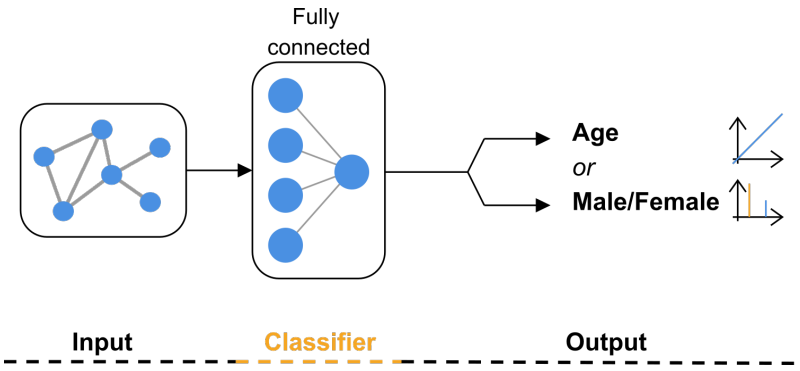


Figure 3.3: The baseline regression model. The adjacency matrix for a brain graph for a single subject forms the input, with a continuous value or class prediction as output in the case of age or sex prediction.

From the definition of a graph in Section 2.1, it follows that graphs are completely node order invariant, and several different ways of listing the connections results in the same graph. This might generally impose a problem since a regression model depends on its input being ordered. In this thesis, a regression model was, however, possible as a baseline model, since the brain graphs from the UK Biobank followed a consistent node ordering, and each input feature of the model was thus always the same connection. Furthermore, the regression model was feasible due to the relatively small size of the brain graphs (21 nodes) compared to many other graphs (such as citation networks consisting of thousands of nodes). For larger graphs, models with one learnable parameter per connection would result in huge models. In that case, models that do not scale with the number of connections must be used, such as GCN-based models.

3.2.2 GCN

The motivation behind using GCN-based models was that they hopefully would be able to extract information from the topological structure of the graphs, rather than simply studying the individual connections as, for instance, Baseline. A GCN-based model, referred to simply as GCN, was therefore developed. An illustration of the GCN model can be seen in Figure 3.4. GCN consisted of three consecutive graph convolutional layers, with propagation rule as defined in equation (2.9), followed by a fully connected output layer. The graph convolutional layers had a Rectified Linear Unit (ReLU) activation function and ten output features. This will be the case for all graph convolutional layers mentioned in the remainder of this thesis, unless otherwise specified. The input to the fully connected output layer consisted of the activations for all three graph convolutional layers, i.e., the activation for all feature maps in all layers, concatenated together. As described in Section 2.3 the activation of layer i contains information about the i:th-order neighbourhood of each node. The inclusion of the activations after each layer in the classifier thus aimed to utilise information of how each node was embedded in a successively larger neighbourhood, which could be beneficial for the predictions.

The inputs to a graph convolutional model generally consists of an adjacency matrix

A and a node feature matrix X. However, the brain graphs in this thesis did not contain any information beyond connections that could be associated with specific nodes. GCN thus used a *featureless* approach, in which the feature matrix was taken to be the identity matrix, $X = I$ [15].

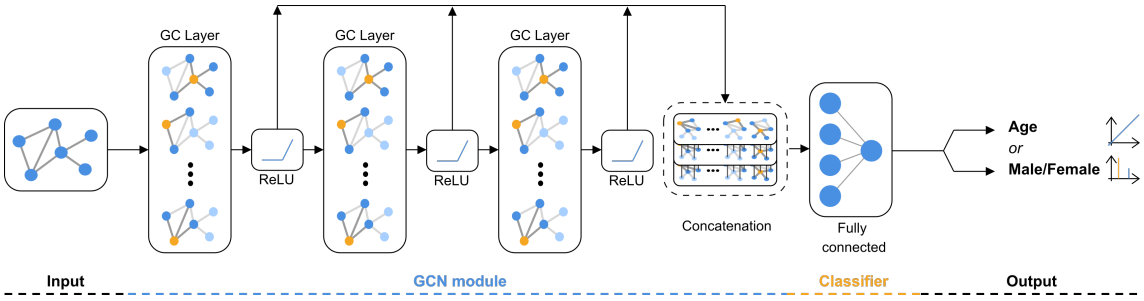


Figure 3.4: The GCN model, which takes a single brain graph as input and outputs a predicted age or sex. The input graph is processed through three graph convolutional layers with ReLU activation, after which the activations for each layer are concatenated together and fed into a fully connected layer.

3.3 Models for node prediction

With the models performing graph predictions introduced, two models used to perform node prediction on a population graph will now be presented. First, however, the procedure of forming population graphs will be described, followed by a batching method that enables models to be trained.

3.3.1 Forming population graphs

As discussed in Section 2.2, the design of the similarity measure is an important decision and should be done with the given application in mind. For our specific application, it is desirable that subjects that have similar fMRI data are connected with edges that have large weights, with the idea that the models will be able to draw upon this information of similarity to yield a better inference. A general construction that fulfils this requirement is the use of a distance metric inverted by a kernel, as described in Equation (2.3). With this construction, and given two subjects and their adjacency matrices A_1 and A_2, we defined the similarity measure $\sigma(A_1, A_2, l)$ as

$$\sigma(A_1, A_2, l) = \exp\left(-\frac{||A_1 - A_2||_F^2}{l||A_1||_F||A_2||_F}\right)\Bigg|_{l=0.5}, \tag{3.1}$$

where $||A_1 - A_2||_F$ is the matrix Frobenius norm of the difference between A_1 and A_2. The Frobenius norm is defined as $||A||_F = \left(\sum_i \sum_j |A_{ij}|^2\right)^{1/2}$. The norm of the difference was weighted with a hyperparameter $l = 0.5$ and the norms of A_1 and A_2, and then fed into a Gaussian kernel. The Gaussian kernel ensured that larger differences between A_1 and A_2 yielded smaller similarity scores $\sigma(A_1, A_2, l)$, and also that $\sigma(A_1, A_2, l) \in [0, 1]$. As desired, subjects that had similar fMRI data, and thus a smaller difference between their adjacency matrices, obtained a larger similarity score and vice versa.

3.3.2 Batches of population graphs

The adjacency matrix for a population graph quickly becomes very large, since the number of edges in the graph grows quadratically with the number of subjects. For example, a population graph with 30 000 subjects requires approximately 7 GB of memory to store. This is a problem, since predicting and training using such a large matrix is time consuming, and the memory consumption can also be problematic. To resolve this problem, and thus in a feasible way be able to train models on population graphs including all the available data, a batching method was developed.

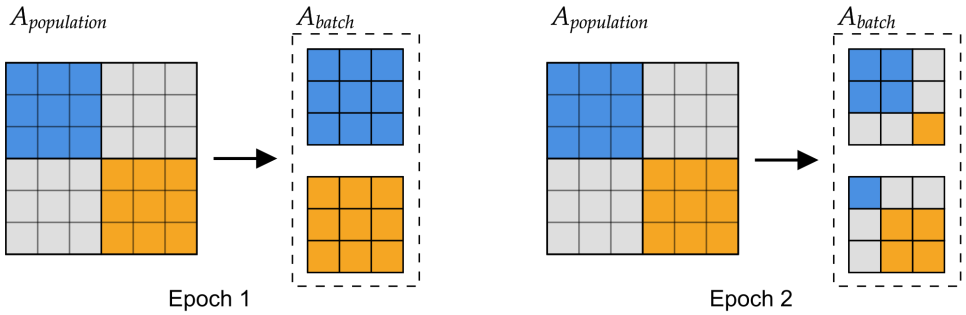


Figure 3.5: An example of the batching approach for splitting a large population graph $A_{population}$ into two smaller population graphs, each corresponding to a batch A_{batch}. In the first epoch, only the connections between the first three (blue) and last three (orange) subjects are included in the batched population graphs. By permuting which subjects are included in each of the smaller population graphs for the second epoch and onwards, all connections not included in the first epoch (gray) will eventually be sampled.

The batching method was based on dividing the population graph into several smaller population graphs. This was done by splitting the data set into several smaller data sets of 100 subjects each, and constructing one population graph for each smaller data set. In practice, this was done by extracting all connections between the 100 subjects from the larger population graph. For an illustration of how this was done, see Figure 3.5. With this batching approach, dividing a population graph of 30000 subjects into, for instance, 300 smaller graphs with 100 nodes each, would require 24 MB of memory, as compared to 7 GB before batching. This is a difference of roughly two orders of magnitude.

As seen in Figure 3.5, many of the connections between subjects are not utilised if the population graph is divided into several smaller graphs. To solve this, the way the subjects were divided into smaller graphs was changed between epochs, as seen in Figure 3.5. By always changing which people were combined in the graphs, all connections were eventually used after enough epochs. The advantage of always changing the subjects that were included in the graphs was that the models could not overfit against a specific graph structure. Thus, the models were forced to learn very general patterns to make predictions for all subjects in the graphs. However, this generalisation might also be a disadvantage, since overfitting on the graph structure may yield higher validation performance, as long as the graph structure remains fixed.

3.3.3 Poptoy

As a first model for predictions on a population graph, the Poptoy model was introduced. The input to Poptoy consisted of a population graph, which was then propagated through five graph convolutional layers. The architecture of Poptoy was similar to that of GCN, but the output layers differed. The output layer of Poptoy was a fully connected layer, and took the activations of all layers and all features for a specific node as input. GCN also had a fully connected output layer, but it took the activations for all layers and all features for all nodes as input. Since the fully connected layer for Poptoy outputs a prediction for a single node, it could be reused for all nodes and thus the number of weights was kept low as the number of nodes in the population graph grew. An illustration of the Poptoy model can be seen in Figure 3.6.

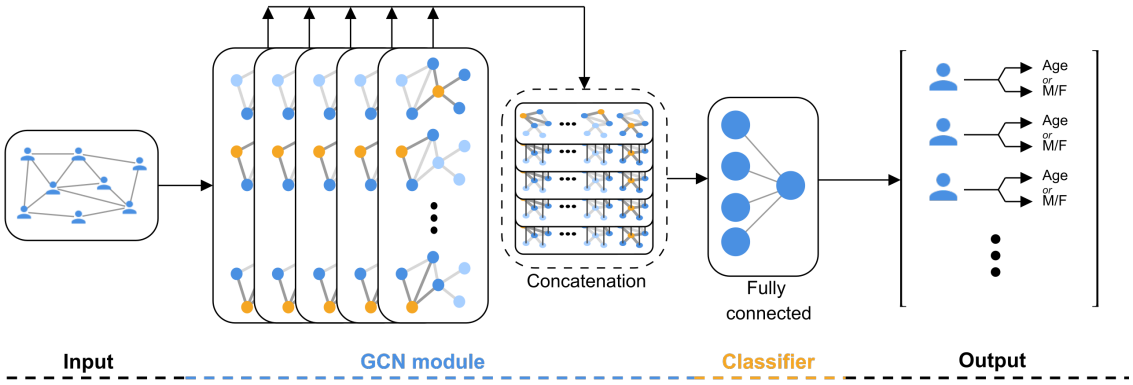


Figure 3.6: The Poptoy model. The input consists of a population graph, which is passed through five graph convolutional layers. The five activations are concatenated together and fed into a fully connected output layer, which outputs a predicted age or sex on a subject-level.

The reason the Poptoy model consisted of five graph convolutional layers instead of three, which GCN had, was that the population graphs became much larger than the individual brain graphs. The individual brain graphs consisted of 21 nodes, which were all more or less connected, and since each graph convolutional layer takes one higher order of neighbours into account, all nodes were to some extent included in the first order neighbourhood. Thus, the need for more graph convolutional layers quickly diminished. For population graphs, the need for considering neighbours farther away in the graph might be much larger, and hence Poptoy utilised more graph convolutional layers than GCN. Five layers were specifically chosen, since adding more layers did not increase performance.

Furthermore, since all subjects in the data set were incorporated in the population graph, the split into validation and training sets had to be handled. To solve this, a set of subjects in the population graph were defined to be the training set, and the rest to be the validation set. The model was then constructed to either do predictions for only the nodes in the training set or the validation set. Then, the training could be performed while only doing predictions on the training set and

vice versa when evaluating.

3.3.4 Popencoder

As a means to incorporate more information about each subject in the population graph, a model referred to as Popencoder was designed. Popencoder is identical to Poptoy in the sense that the population graph was propagated through five graph convolutional layers, after which a prediction for each node was made using a fully connected layer. The difference is that in Popencoder, features for each node were introduced. The features were based on the adjacency matrices of each individual subject, but to compress the dimensionality of the feature space, these matrices were encoded into a lower dimensional space. The encoder consisted of two graph convolutional layers, followed by a fully connected layer with either two softmax activated neurons or one linearly activated neuron, in the case of sex or age prediction, respectively. For an illustration of Popencoder, see Figure 3.7.

A heuristic explanation of why the features introduced to the population graph would help, is that the encoder can make an initial prediction of the age or sex for each individual subject. Then, by propagating this information through the population graph, the information of similarities to other subjects in the data set might improve on the predictions. In that case, one could argue that the initial prediction might be done as a preprocessing step, for example via a prediction with another model. The encoding is, however, viewed as a trainable part of the model to allow for more abstract and advantageous embeddings to be learnt during model training.

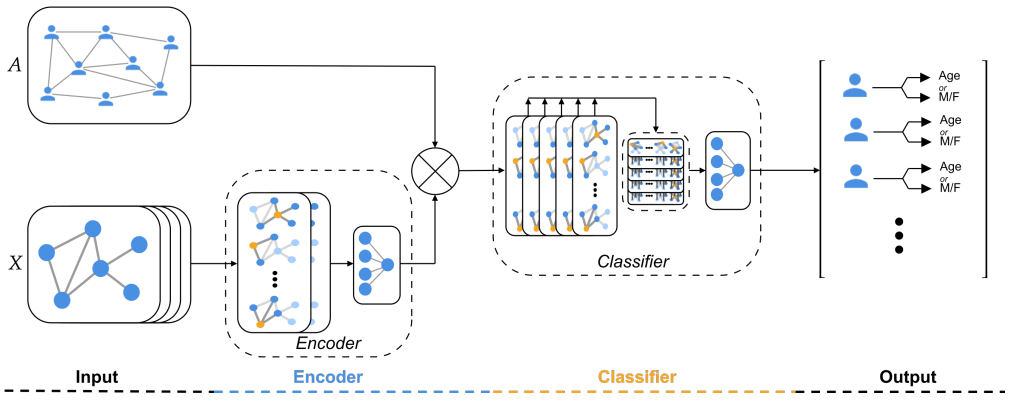


Figure 3.7: The Popencoder model, which takes two inputs: a population graph A, and the brain graphs for all subjects in the population graph, X. The brain graphs are passed through an encoder, consisting of two graph convolutional layers followed by a fully connected layer. The encoded brain graphs and the population graph are then fed into a classifier, which consists of five graph convolutional layers followed by a concatenation and a fully connected layer. The output is a predicted age or sex on a subject-level.

3.4 Model explainability through saliency mapping

To determine what functional networks in the brain are related to sex and age, model analysis, in the form of analysing what the models have learned, is an essential part of the work. Specifically, using saliency mapping techniques to analyse which nodes in the brain graph are important for making predictions will, by extension, give information on what functional networks are important, due to each node representing a functional network.

Two methods for model analysis were used: a naive approach based on node removal, and a more sophisticated method for node masking based on the Zorro algorithm described in Section 2.4. Note that these are not methods that open up the black-box of neural networks per say (by e.g. parameter analysis). These methods analyse the networks from an outer perspective, which is beneficial as it poses no assumption on the model architecture, only in- and output data. Thus, the same analysis may be performed for different models, making it beneficial for validation and result comparison.

3.4.1 Naive node removal

A simple method to obtain a saliency map of which nodes are important for predictions, is a node removal method. This particular method is self-composed, and consisted of simply removing a specific node for every subject in the data set, after which a new model was retrained. By comparing the predictive performance of the retrained model with a reference model trained on data with no nodes removed, an indication of the importance of that node could be obtained. This is based on the assumption that a large loss in performance means vital information for the predictions had been removed, which was interpreted to be indicative of the importance of that node. The method was then repeated for all nodes in order to obtain a measure of importance for each node.

3.4.2 Zorro

The Zorro algorithm, described in Section 2.4, was developed for saliency mapping of models with graph features as input. It thus needed to be modified to be applied to the models presented in this thesis. The need for modification arose since Baseline and GCN are completely featureless approaches, that only takes an adjacency matrix A as input. Therefore, it was not possible to introduce noise to the feature matrix, as done in the original method. Instead, it had to be done on A. The reintroduction of nodes in A, in the following referred to as unmasking, could have been done in two ways; either on a connection level where entries in A are unmasked, or on a nodal level where whole rows and columns in A are unmasked. We were primarily interested in which nodes were important, so the latter was used. Unmasking whole nodes instead of connections also yielded computational benefits, since the number of nodes was less than the number of connections. As whole nodes were unmasked, the

explanation in the modified Zorro method only contained a set of nodes, $\mathcal{S} = \{V\}$. The masked adjacency matrix was then given by

$$B_S = A \odot S + Z \odot (1 - S), \quad Z \sim \mathcal{N}, \tag{3.2}$$

where S is the masking matrix for the explanation $\mathcal{S}$, 1 is a matrix of ones and $\odot$ represents element-wise multiplication. The noise Z was drawn from a Gaussian distribution with mean and standard deviation given by the entries of A over all subjects in the data set. The model prediction on the masked adjacency matrix $\Phi(B_S)$ was considered correct for sex prediction if $\Phi(A) = \Phi(B_S)$, and for age prediction if

$$|\Phi(A) - \Phi(B_S)| < t, \tag{3.3}$$

for some tolerance t, since age is a continuous variable. The fidelity was calculated in the same way as in the original algorithm, and an explanation $\mathcal{S}$ for an individual subject was still accepted if the fidelity of $\mathcal{S}$ was higher than τ.

Lastly, since the algorithm yielded which nodes were important for the prediction of an individual subject, the procedure was repeated for several subjects to get a sense of which nodes were generally important. To evaluate the importance of each node, an importance score $\mathcal{I}$ was introduced as the number of explanations $\mathcal{S}$ a node was included in, divided by the total number of subjects. An importance score of $\mathcal{I} = 1$ indicates that the node is considered important for the prediction of all subjects, and a score of $\mathcal{I} = 0$ for none.

4. Results

This chapter consists of two main parts. In the first part of the chapter, the results from training models for predicting sex and brain age from brain graphs will be presented. The models were first evaluated using ten-fold cross validation on the training/validation set. Then, a final model was trained and evaluated on the test set. In the second part of the chapter, the saliency maps for these final models will be presented, with the aim of gaining insight into what functional brain networks in the data are related to age and sex differences.

To reliably evaluate the models, the data set was split into a training/validation set and a test set. These data sets where then undersampled to make them unbiased, which yielded training/validation sets of size 30000 and 20000, and test sets of size 4678 and 3148 for sex and age prediction, respectively. For age prediction, an unbiased data set corresponds to a uniform distribution of ages. Model training and evaluation was implemented in `Python` [29] using `Keras` [30] with the `TensorFlow` [31] backend. Furthermore, all models where trained using the `Adam` [32] optimizer.

4.1 Evaluating model performance

The model performance for both sex and age prediction is presented in the form of the loss and two additional metrics. For sex classification, these metrics are Binary Cross Entropy (BCE) (loss), accuracy and Matthews Correlation Coefficient (MCC). For age regression, they are Mean Squared Error (MSE) (loss), Mean Absolute Error (MAE) and Pearson correlation coefficient (r). MCC is defined identically to r, but refers to the performance of a binary classifier. An MCC of +1 thus indicates a perfect prediction, -1 all miss-predicted and 0 randomly guessing.

4.1.1 Sex

To evaluate the model performance for sex classification, the three metrics calculated from cross-validation on the training/validation set are presented in Table 4.1. All three metrics indicate that Baseline, GCN and Popencoder have comparable performance, with an accuracy around 79 %. It is also clear that these three models significantly outperform Poptoy regardless of metric. The performance of the final models, evaluated on an external test set, is presented in Table 4.2. The final models' performance are in line with the cross-validation results in Table 4.1, where all metrics are within, or around, one standard deviation. Thus, the performance of the models generalises to external data.

The results presented in Table 4.1 and Table 4.2 have several interesting implications.

Table 4.1: Binary Cross Entropy (BCE), accuracy (in %) and Matthews Correlation Coefficient (MCC) for each of the four models, evaluated using ten-fold cross validation, with the mean and standard deviation calculated over the ten folds.

	BCE	Accuracy	MCC
Baseline	0.450 ± 0.008	79.0 ± 0.6	0.58 ± 0.01
GCN	0.45 ± 0.02	79.2 ± 0.9	0.59 ± 0.02
Poptoy	0.674 ± 0.006	58 ± 1	0.17 ± 0.02
Popencoder	0.446 ± 0.009	79.3 ± 0.6	0.59 ± 0.01

Table 4.2: Binary Cross Entropy (BCE), accuracy (in %) and Matthews Correlation Coefficient (MCC) for each of the four models evaluated on the test set.

	BCE	Accuracy	MCC
Baseline	0.444	79.5	0.59
GCN	0.43	80.2	0.60
Poptoy	0.678	57.4	0.16
Popencoder	0.450	79.1	0.59

Observe that there is basically no performance difference between Baseline, GCN and Popencoder, implying that increasing model complexity beyond Baseline is not necessary for classifying sex. Furthermore, the main difference between Popencoder and Baseline and GCN is that Popencoder takes a population graph as input, in addition to the brain graphs for all subjects. One explanation of why including the population graph does not improve performance could be that our chosen similarity measure does not introduce any extra information relevant for sex classification. Since the similarity measure is calculated from each subject's brain graph, it is possible that the information encoded by the similarity measure is only a subset of all the information contained in the individual graphs. Another explanation is that the similarity measure adds some extra information, but that the amount of information is small. In the case of either explanation, the similarity measure does not introduce enough information to improve the prediction of a subject's sex. That the similarity measure is not very informative is further supported by the poor performance of Poptoy, since Poptoy bases its prediction solely on the population graph. However, Poptoy still performs better than random guessing, indicating that some valuable information resides in the population graph.

4.1.2 Age

To evaluate the model performance for age regression, the three metrics calculated from cross-validation are presented in Table 4.3. From Table 4.3, the three metrics clearly indicate that, just as for sex classification, Baseline, GCN and Popencoder all perform within the uncertainty of each other, whilst Poptoy performs significantly worse. Table 4.4 presents the performance of the final models trained on all data and evaluated on an external test set. The results are within one standard deviation of the results in Table 4.3, which indicates that the models generalise well to external data. One possible exception is Popencoder, which performs slightly better on the test set than during cross-validation, however, the difference in performance is small.

Table 4.3: Mean Squared Error (MSE), Mean Absolute Error (MAE) and Pearson correlation coefficient (r) for each of the four models, evaluated using ten-fold cross validation, with the mean and standard deviation calculated over the ten folds.

	MSE [years^2]	MAE [years]	r
Baseline	52 ± 1	5.9 ± 0.1	0.52 ± 0.01
GCN	53 ± 1	5.96 ± 0.09	0.52 ± 0.01
Poptoy	71 ± 1	7.24 ± 0.07	0.11 ± 0.01
Popencoder	53 ± 1	5.93 ± 0.09	0.52 ± 0.02

The age prediction results for the different models are in line with the two main conclusions for sex classification in the previous section. Firstly, Baseline, GCN and Popencoder perform similarly which indicates that increasing the model complexity beyond Baseline is unnecessary. Secondly, the poor performance of Poptoy compared to the other models, combined with Popencoder not performing better than Baseline and GCN, indicates that the similarity measure is not very informative with regards to age either.

Table 4.4: Mean Squared Error (MSE), Mean Absolute Error (MAE) and Pearson correlation coefficient (r) for each of the four models evaluated on the test set.

	MSE [years^2]	MAE [years]	r
Baseline	51	5.9	0.53
GCN	51	5.9	0.53
Poptoy	71	7.18	0.09
Popencoder	50	5.7	0.55

To set the poor performance of Poptoy into perspective, consider a naive age prediction model that always outputs the average age of the population when predicting a subject's age. The performance of such a model on the test set would yield an MSE of 71 years^2 and an MAE of 7.3 years. The correlation for constant prediction with the actual ages is undefined. Comparing this with the results for Poptoy, one can conclude that Poptoy only performs slightly better, demonstrating Poptoy's poor performance.

Turning to the results on the test set for the three other models, all have an MSE loss of around 50–51 years^2 , an MAE of 5.7–5.9 years and a correlation of around 0.53–0.55. The fact that the correlation is much lower than 1 indicates that the prediction is not perfect. In fact, it turns out that all these models are somewhat biased in predicting the mean of the age distribution in the data set. This can be seen clearly in Figure 4.1, where a prediction with Baseline on the test set is performed. The figure suggests that the model to some extent can determine if a subject is old or young, but that the error in the age prediction is much larger for younger and older subjects than for subjects close to the mean age. This is a form of underfitting, indicating either that the model is not complex or general enough to better fit the data, or that the data does not contain enough information for a better prediction of age.

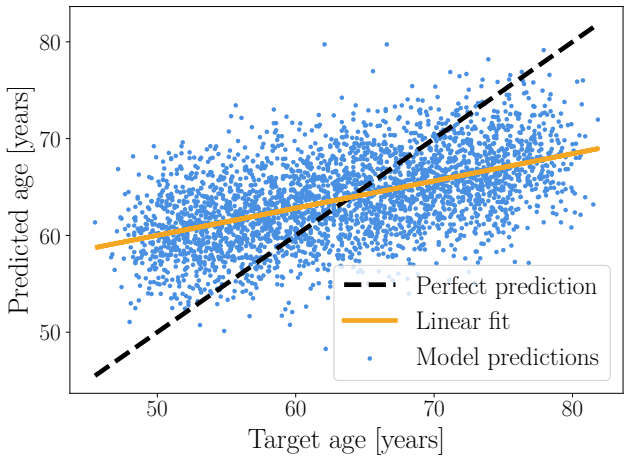


Figure 4.1: A plot of the Baseline predicted ages for the subjects in the test set, versus the target ages. The dots are the predictions for individual subjects, the orange line is a linear fit to the individual predictions, and the black dashed line is what the perfect predictions (no error) would look like. Note that the error in the prediction is the largest for the youngest and the oldest subjects.

4.2 Saliency maps of functional brain networks

The saliency mapping consisted of two methods: naive node removal and Zorro; both applied for sex and age prediction on Baseline and GCN. Poptoy was not analysed because of its low performance, and Popencoder because of its high complexity. Including Popencoder would probably also be redundant, because of its similar performance to Baseline and GCN.

Naive node removal was performed with the difference in loss (ΔBCE for sex prediction and ΔMSE for age prediction) evaluated on the test set. The naive analysis was repeated ten times with different model initialisations to estimate the impact of model uncertainty on node importance. Zorro was performed only for the final models presented in Section 4.1, and was run with a fidelity threshold of $\tau = 0.9$. The tolerance for an age prediction to be considered correct was set to be similar to the MAE of the final models, at $t = 6$ years. The results for all subjects in the test set were aggregated into groups of 200 subjects, and the mean importance score for each node in each group was calculated. Then, the average and standard deviation over the importance scores in each group was calculated. The standard deviations of the Zorro results over these sub-groups represent the uncertainty in the importance for each node when varying subjects. Ideally, Zorro should also be evaluated over different model intialisations. However, it would require repeating Zorro evaluated on all subjects in the test set for several different models. Due to the computational complexity of Zorro, this was not possible because of time constraints. A smaller investigation into the model uncertainty for Zorro was performed for a few subjects, presented in Appendix B, from which it was concluded that the model uncertainty is similar, but generally smaller than the subject uncertainty.

4.2.1 Sex

As a measure of node importance, ΔBCE when performing naive node removal for Baseline and GCN is presented in Figure 4.2.

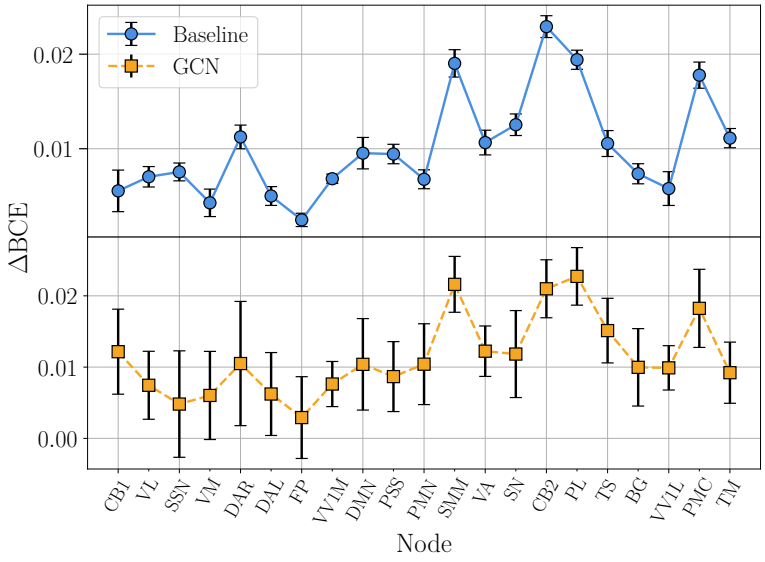


Figure 4.2: Results from performing the naive node removal analysis for Baseline and GCN, for sex prediction. The analysis was repeated for ten different model initialisation, over which the dots and error bars in the figures represent the mean and standard deviation, respectively.

From the figures it is observed that the effect of removing a node for both models is on the order of $\Delta\text{BCE} \sim 0.01$, which is small compared to the absolute performance of the reference models, $\text{BCE} \approx 0.45$. It is also observed that removing any node has a negative effect on performance, $\Delta\text{BCE} > 0$. These two observations indicate two things. Firstly, no node seems to be crucial for the prediction, since the change in performance is small for all nodes. Secondly, all nodes are to some extent important, since removing any node has a negative impact on the performance. Despite the small absolute change in performance, there are clear differences in the relative importance between nodes. The results for both models indicates that SMM, CB2, PL and PMC are more important than the other nodes. Further, observe that the uncertainties are much larger for GCN than for Baseline. This could be due to GCN not converging to the same degree as Baseline, either due to overfitting, early stopping or because of it being a more complex model. The two models, however, give roughly the same result, even if the uncertainties for GCN are large.

The result from the Zorro analysis for Baseline and GCN is presented in Figure 4.3. The uncertainties in the result for both models are small, but note that they do not include potential uncertainties regarding the model variability, as previously discussed. Generally, it is clear that the results for Baseline and GCN are to a large extent in agreement: for both models SMM is pointed out to be the most important node. Other nodes that may be regarded to be somewhat important

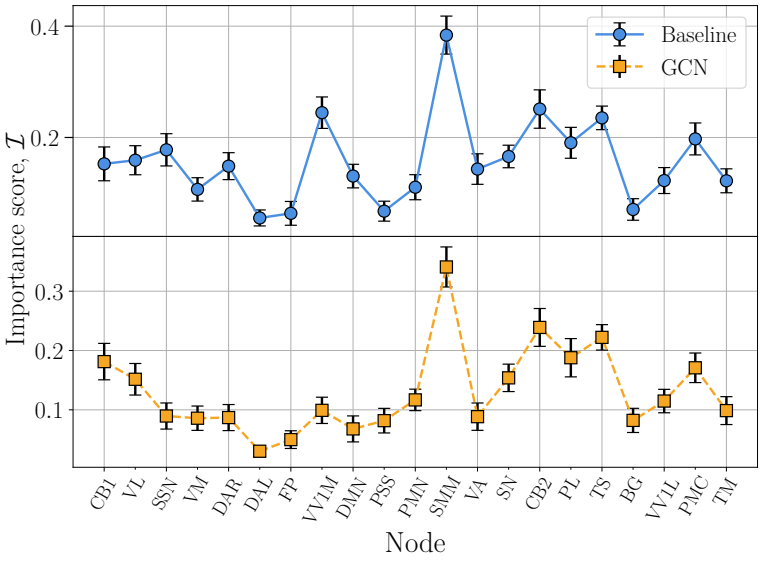


Figure 4.3: Zorro analysis results for the Baseline and GCN models. The analysis results were grouped into 23 different groups of 200 subjects for each model, to yield a mean and standard deviation in importance for each node.

according to both models are CB2, PL, TS and PMC. It is also interesting to note that VV1M seems to be more important when analysing Baseline than GCN. One possible interpretation could be that this is an artefact of the method, but this inconsistency will be discussed more in-depth in Chapter 5.

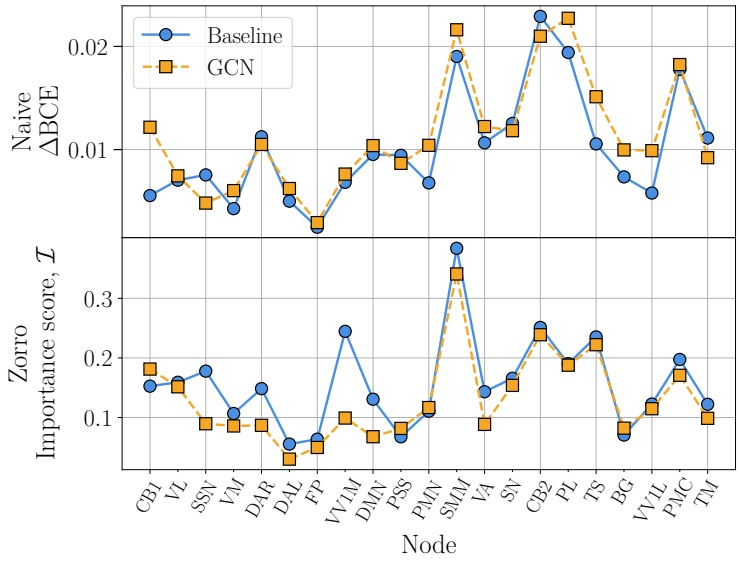


Figure 4.4: Comparison of the results for naive node removal and Zorro for sex prediction, from Figure 4.2 and Figure 4.3, respectively. Note that the error bars are omitted for visibility.

To more easily compare the result of naive node removal and Zorro for both models, Figure 4.4 is presented, without error bars for visibility reasons. The comparison of the result reveal that the methods are generally in agreement, but some clear differences exist. For instance, both methods agree that SMM is important but disagree on the degree of importance for CB2 and PL. By comparing the methods and models, SMM seems to be most important node for classifying sex, and CB2, PL, TS and PMC may to some extent be important.

4.2.2 Age

As a measure of node importance, ΔMSE when performing naive node removal for Baseline and GCN is presented in Figure 4.5.

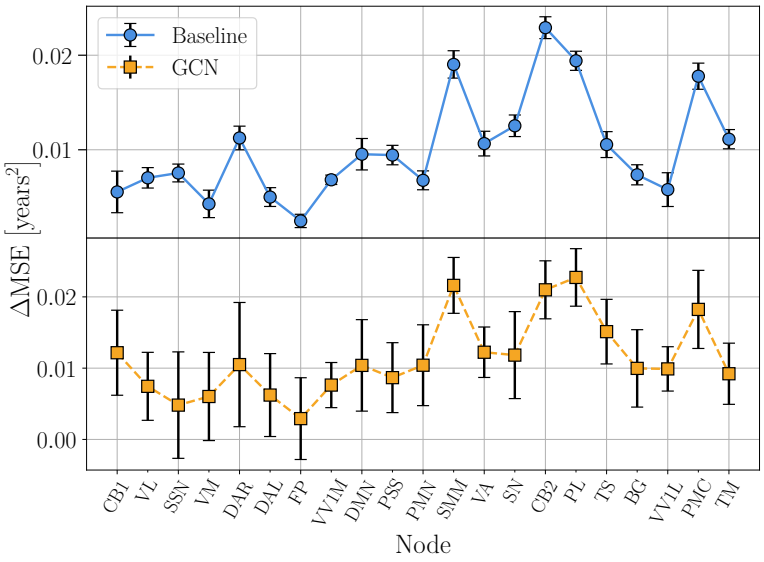


Figure 4.5: Results from performing the naive node removal analysis for Baseline and GCN, for age prediction. The analysis was repeated for ten different model initialisation, over which the dots and error bars in the figures represent the mean and standard deviation, respectively.

The general change in loss is observed to be small, $\Delta\text{MSE} \sim 1\,\text{years}^2$ relative to $\text{MSE} \sim 50\,\text{years}^2$. As was the case for sex prediction, this indicates that no node is crucial and that all nodes are to some extent important for predicting age. Furthermore, the model uncertainty is larger for GCN than for Baseline in the case of age prediction as well. This could be due to GCN not fully converging, as discussed in Section 4.2.1. Observe that the most important nodes for both models are SMM, CB2 and PMC, where CB1 is additionally important for Baseline. Overall, the models are generally in agreement with some small exceptions.

The result from the Zorro analysis for Baseline and GCN is presented in Figure 4.6. Observe that the algorithm finds CB1, DMN, SMM, SN, CB2 and PL important for Baseline, and SMM and CB2 important for GCN. Another observation is that generally the importance scores for Baseline are higher than for GCN, especially for

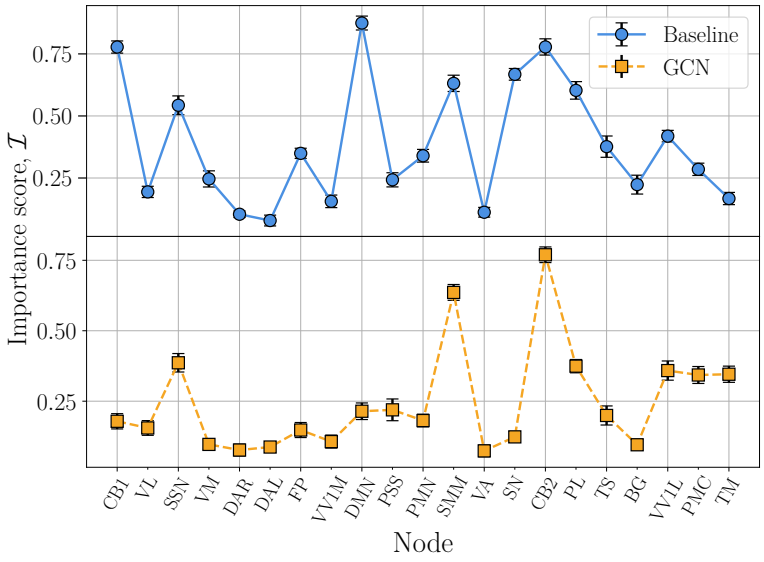


Figure 4.6: Zorro analysis results for the Baseline and GCN models for age prediction. The analysis results were grouped into 15 different groups of 200 subjects for each model, to yield a mean and standard deviation in importance for each node.

CB1, DMN and SN. This means that more nodes are required in the explanations for Baseline than for GCN in order to reach a fidelity of $\tau = 0.9$.

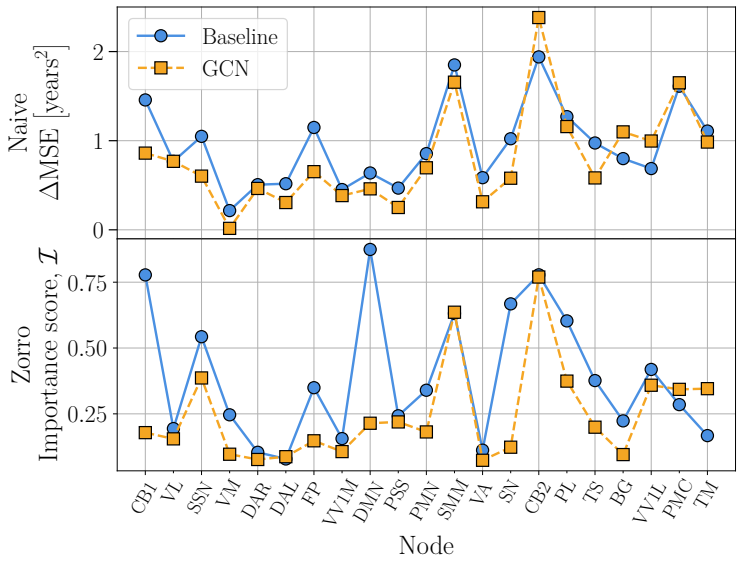


Figure 4.7: Comparison of the results for naive node removal and Zorro for age prediction, in Figure 4.5 and Figure 4.6 respectively. Note that the error bars are omitted for visibility.

In order to more easily compare the result for the naive node removal and Zorro for both models, Figure 4.7 is presented. By comparing the results for both methods and models, SMM and CB2 stand out, since they are deemed important in all

cases. In addition to SMM and CB2, other nodes that might be important, but not consistently for all methods and models, are CB1, DMN, SN, PL and PMC. Most of these are nodes that are only considered important by Zorro for the Baseline model. However, since they are only deemed important in one out of the four cases, it might suggest that they are artefacts of the method or model. There are also some differences in the results for naive node removal compared with Zorro. For example, PMC may be deemed more important by the naive method. However, the differences between the two analysis methods seem to be smaller than the discrepancy between Baseline and GCN for Zorro. These differences and discrepancies will be discussed in more detail in Chapter 5.

5. Discussion

Our results indicate two main findings with regards to the model performance. Firstly, there seems to be no benefit in using more complex models than the Baseline regression model, and secondly, the similarity measure does not seem to introduce enough information to improve the predictions of sex and brain age. A possible explanation as to why there is no benefit in using more complex models over Baseline could be that the data is too low dimensional. The simple Baseline model is perhaps able to extract all relevant information from the 21-node graphs, and hence the more complex models cannot improve upon the performance of Baseline. Increasing the dimensionality of the fMRI data could thus potentially increase the performance for the models presented in this thesis. That the performance could be improved by using higher dimensional fMRI data is further supported by similar works in the field, which reach higher predictive performance using similar models, but with larger graphs. For sex classification, [13] and [26] reaches classification accuracies of 84% and 88%, using fMRI brain graphs with 400 and 55 nodes, respectively. This is likely the case for age prediction as well, where for example [12] obtains an MSE of 27.5 years^2, [2] an MAE of 4.7 years and [33] an MAE of 2.14 years, all using high-dimensional MRI data. Note though that the results in [2], [12], [33] are not directly comparable to ours, due to the use of MRI and not fMRI data.

The limitation imposed by the low data dimensionality motivated the introduction of the population graph, as an attempt at improving the performance of the GCN models. The heuristic explanation for why population graphs could be effective is that an initial prediction of sex or age could be corrected by considering more or less complex structures of similarities in the data set. That this would improve performance is based on the assumption that people with high similarity scores would have approximately the same age or sex. The absence of a performance increase for Poptoy and Popencoder suggests that the similarity measure cannot indicate if two subjects have similar age or sex, and the need for a better similarity measure is evident. Other approaches for similarity measures are presented in [12] and [34], which use non-imaging data, such as education and cognitive abilities, and a graph theoretical similarity measure, respectively. Another way to improve the similarity measure is to include more domain knowledge. Domain knowledge could for example be knowledge about which functional brain networks are related to sex and age, which is part of the analysis in this thesis, but other types of domain knowledge could also be useful. Furthermore, an informative similarity measure could be learnt using machine learning, which would enable complex and abstract measures, that perhaps would be hard to find even for domain experts [35]. Investigating more sophisticated similarity measures was however not possible in this study, and is left as future research.

The result from the saliency mapping state that all functional brain networks are to some extent important, and that no single node is crucial for predicting sex or age. The fact that all nodes are important for the predictions was expected, since it is reasonable that the whole brain is affected during e.g. ageing. The results also show significant variations in how important different nodes are. Since the data used in this thesis consist of connections between functional brain networks, the interpretation of a node being important is that its connections to all other functional networks are important for predicting sex or brain age. That some nodes are more important is thus interpreted that some functional networks, or at least parts of the functional networks, are to a higher degree related to differences in sex and age. Furthermore, the two analysis methods are generally in agreement, but there are some inconsistencies. More precisely, the result from Zorro when analysing Baseline and GCN are inconsistent with each other, where Baseline obtains higher importance scores than GCN for CB1, DMN and SN for age, and VV1M for sex. These inconsistencies are very interesting, and to be able to conclude which nodes are most important they need to be discussed.

The inconsistencies may be interpreted as that different functional networks are regarded as important by Baseline compared to GCN. Even if this is possible, it is deemed unlikely, because of the similar performance of the two models. To explain the inconsistencies if both models find the same functional networks to be important, it is imperative to realise that the obtained results are only valid for $\tau = 0.9$. Changing τ may affect the results for both Baseline and GCN, where increasing/decreasing τ would lead to more/fewer nodes being added to the explanations. This enables two possible interpretations: CB1, DMN, SN and VV1M are either not important for age and sex predictions, or they are important. In the case of the nodes not being important, τ could have been selected too high, meaning the nodes are included to recreate specific predictions rather than being indicative of sex or age. In the case of the nodes being important, it is possible that the nodes will be added to the explanations for GCN for a slightly higher value of τ. In this scenario, the nodes could be regarded as less important than SMM and CB2, since they are added last to the explanations. However, to be certain what actually causes the inconsistencies in Zorro, other values of τ would need to be investigated, which was not done due to time limitations.

Having discussed the inconsistencies regarding the Zorro results, the small but nevertheless present differences between the naive node removal method and Zorro need to be addressed. These differences may perhaps be reduced by another choice of τ, since changing τ would affect the result of Zorro. Even so, it is still important to remember that the two methods answer slightly different questions. Naive node removal answers the question which removed node affect the performance of the models the most, whilst Zorro answers the question which node is most important to unmask for a prediction to be as close to the original as possible. Even if the two questions seem to be very similar the results should not be expected to be identical. However, both questions still indicate how important nodes are for the predictions, and using two different methods can thus strengthen the credibility of the results.

Extending the analysis to include more methods could thus be beneficial, but this is left as future work.

With the two saliency mapping techniques discussed, we can now re-examine the results in this new light. Comparing the methods strengthens the interpretation that VV1M for sex and CB1, DMN and SN for age are not among the most important nodes. Furthermore, comparison shows that SMM is most important for sex classification followed by CB2 and PL. For age prediction, comparing the two methods for both models clearly indicates that SMM and CB2 are the most important nodes and that PMC also may be somewhat important. Interestingly, SMM and CB2 are deemed important for both sex and age prediction. This overlap may be expected, since sex and age are covariates [36], in the sense that changes in functional connectivity of the brain that are related to age may also be related to sex. In this work we have focused on the effects of sex and ageing in general, without controlling for the other. This may be a limitation in the sense that the analysis e.g., compounds the effects of sex-related differences in functional connectivity when studying ageing, and vice versa. More in-depth analysis, in which sex and age are controlled for, would thus be a suitable direction for future work.

Comparing our results with the literature, we find both similarities and inconsistencies. For sex prediction, [26] performed a saliency mapping using similar machine learning models as in this thesis, and arrived at the DMN and Sensorimotor Network (SMN) being the most important for sex prediction. In [13], the DMN was found to be important. For age prediction, Song et al. [37] found the DMN and SMN to undergo relatively large changes in connectivity with age. Tomasi et al. [38] studied the connections between networks and found, among other networks, that the DMN, the Dorsal Attention Network (DAN), the Somatosensory Network (SSN) and cerebellum exhibited pronounced changes in functional connectivity with age. A machine learning approach using SVMs was used in Meier et al. [39], where the DMN, SMN and cingulo-opercular networks were found to be related to age. Most of these works are to some extent in agreement with our results of which networks are important for predicting sex and brain age. The SMN, which includes the SMM that we find important, is often found to be among the most important networks for both sex and age, and additionally, the cerebellum is found to be related to age in [38]. However, one major inconsistency is that all of the studied articles find the DMN to be among the most important networks.

One possible explanation for the inconsistencies regarding the DMN may be the dimensionality of the data. In several of the studies mentioned, higher dimensional data has been used. The DMN and other networks thus consisted of several nodes, but in this thesis they were represented as single nodes. It is possible that the useful information for sex or age prediction resides in the connections within the DMN or in how individual parts of the DMN are connected to other networks. This could explain both the fact that we obtain slightly lower performance than for example [13], [26], and why DMN is not considered to be among the most important networks in our thesis. However, it is still important to remember that we find all networks

to be important, including DMN. Several of the studies which have observed age related changes in DMN [37]–[39] have no way of ranking the networks, and thus our observation of SMM and CB2 being more important than DMN does not need to be an inconsistency. In [13], [26] they have the ability to rank importance of nodes for sex classification, and they find DMN to be one of the most important, while not even mentioning the cerebellum. An explanation for this disagreement, besides the data dimensionality, could be the different methods used to perform the saliency mapping. Just as for the naive node masking method and Zorro, the methods in [13], [26] might answer different questions than our methods, and thus yield different results. This highlights the complexity of understanding machine learning models, and simply stating that a node is important might be insufficient. To be sure a specific network is related to sex or age, the important question to answer is why this network is important for the prediction.

Since the saliency mapping methods used in this thesis are based on investigating the machine learning models, the significance of the analysis results is limited by the model performance. The underlying assumption for the analysis to be meaningful is that the models can extract relevant information for predicting sex and brain age from the data. Thus, any limitations on the model performance is by extension a limitation on the analysis. In this sense, the low dimensionality of the data used to train the machine learning models hinders the analysis. Because of this, using higher dimensional data is crucial for improving the saliency maps. However, using higher dimensional data can make the analysis computationally more challenging, and interpretation of the results more difficult. From this perspective, one could argue that a decrease in model performance is worth it for the sake of interpretability of the results. Specifically, the 21-node data used in this thesis has the benefit of each node roughly corresponding to a functional brain network, which greatly simplifies the interpretation of the saliency maps. Another benefit of using the 21-node data is that, as our results indicate, the use of more complex models is not necessary. Using simpler models, such as the Baseline regression model, makes the models themselves more explainable. Explainability is especially important for potential clinical applications, where machine learning is applied to patients. Thus, there exists a clear trade-off between model performance and explainability. Finding the sweet spot between these two is key to further gain insight into how functional brain networks are related to sex and ageing using machine learning.

6. Conclusion and Outlook

In this study, accurate models based on GCNs for sex and brain age prediction using RS-fMRI data have been developed and analysed, with the goal of identifying functional brain networks that are related to sex and age. Three of the four studied models achieved comparable performance for both prediction tasks, with an accuracy of up to 79 % for sex classification a mean absolute error as low as 5.9 years for age prediction. From the saliency maps it was concluded that all functional brain networks are important for sex and brain age prediction, but that some are more important than others. Specifically, the most important network was the Somatomotor Medial (SMM) for both sex and age prediction, and additionally, the cerebellum was found to be important for age prediction. The main limitation of the work was found to be the low resolution of the RS-fMRI data used in this thesis.

There are several improvements that could be considered in future works. One possible improvement is to study the use of more high dimensional fMRI data. Another improvement is the inclusion of other data modalities, such as using both structural MRI and fMRI data. It is possible that such combined approaches could improve model performance. Different model types, not based on fully connected neural networks or GCNs, could also be investigated. An example of such model types are various gradient-tree boosting algorithms, which have previously been applied to predicting brain age [6].

Studying different kinds of similarity measures is also a possible direction for future work. As previously mentioned, non-imaging data as in [12] can be used. This could increase model performance, but at the risk of introducing confounding variables that might make the analysis more difficult. If only imaging data is considered, as in this thesis, there are still several interesting approaches. One approach is to use graph-based similarity measures, such as in [34]. Another approach is to train a machine learning model to predict the similarity between subjects. The model could thus be seen to compress the imaging data of two subjects into a single number. Finally, domain knowledge can also be used to design the similarity measure. For instance, a similarity measure based on comparing the functional connectivity could be improved by including information on the importance of nodes, e.g., by weighting each node's connections by its importance.

Another direction for future work is to investigate other saliency mapping techniques, but also more general methods for explainability in AI. Both naive node removal and Zorro, used in this thesis, perform input-output analysis of the models. To further investigate the explainability of the models, one could consider using approaches that "open-up" the neural networks. In Yuan et al. [40] a review over

possible analysis methods for GCNs are presented. One example is Grad-CAM [41], a method originally developed for analysing the gradients within CNNs to draw conclusion on the importance of each pixel in the input image for predicting a certain class. Grad-CAM has been generalised to GCNs, and was for instance used in both [13] and [26] to study the importance of functional brain networks for sex classification. We considered using Grad-CAM in this work, but were not able to due to time limitations.

This thesis has mainly focused on gaining insight into which functional brain networks are important for predicting sex and brain age, without approaching the question of why they are important. To approach this question, more sophisticated methods for explainability could be used in order to gain a deeper neuroscientific understanding of the results. Explainability can be seen as the bridge that gaps the intersection between machine learning and neuroscience, and could possibly enable many interesting investigations in the years to come.

References

[1] J. Richiardi, S. Achard, H. Bunke, and D. Van De Ville, "Machine learning with brain graphs: Predictive modeling approaches for functional imaging in systems neuroscience," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 58–70, 2013, ISSN: 10535888. DOI: 10.1109/MSP.2012.2233865.

[2] N. Amoroso, M. L. Rocca, L. Bellantuono, *et al.*, "Deep learning and multiplex networks for accurate modeling of brain age," *Frontiers in Aging Neuroscience*, vol. 11, no. 5, 2019, ISSN: 16634365. DOI: 10.3389/fnagi.2019.00115.

[3] N. Amoroso, M. L. Rocca, S. Bruno, *et al.*, "Multiplex networks for early diagnosis of Alzheimer's disease," *Frontiers in Aging Neuroscience*, vol. 10, 2019-11, ISSN: 16634365. DOI: 10.3389/fnagi.2018.00365.

[4] M. Y. Chan, D. C. Park, N. K. Savalia, S. E. Petersen, and G. S. Wig, "Decreased segregation of brain systems across the healthy adult lifespan," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 46, E4997–E5006, 2014-11, ISSN: 10916490. DOI: 10.1073/pnas.1415122111.

[5] M. Mijalkov, E. Kakaei, J. B. Pereira, E. Westman, and G. Volpe, "BRAPH: A graph theory software for the analysis of brain connectivity," *PLoS ONE*, vol. 12, no. 8, 2017-08, ISSN: 19326203. DOI: 10.1371/journal.pone.0178798.

[6] T. Kaufmann, D. van der Meer, N. T. Doan, *et al.*, "Common brain disorders are associated with heritable patterns of apparent aging of the brain," *Nature Neuroscience*, vol. 22, no. 10, pp. 1617–1623, 2019-10, ISSN: 15461726. DOI: 10.1038/s41593-019-0471-7.

[7] O. Sporns, "Structure and function of complex brain networks," *Dialogues in Clinical Neuroscience*, vol. 15, no. 3, pp. 247–262, 2013, ISSN: 12948322. DOI: 10.31887/dcns.2013.15.3/osporns.

[8] G. V. Hirsch, C. M. Bauer, and L. B. Merabet, "Using structural and functional brain imaging to uncover how the brain adapts to blindness," *Journal of Psychiatry and Brain Functions*, vol. 2, no. 1, p. 7, 2015. DOI: 10.7243/2055-3447-2-7.

[9] P. N. Alves, C. Foulon, V. Karolis, *et al.*, "An improved neuroanatomical model of the default-mode network reconciles previous neuroimaging and neuropathological findings," *Communications Biology*, vol. 2, no. 1, 2019-12, ISSN: 23993642. DOI: 10.1038/s42003-019-0611-3.

[10] C. Grady, "The cognitive neuroscience of ageing," *Nature Reviews Neuroscience*, vol. 13, no. 7, pp. 491–505, 2012-07, ISSN: 1471003X. DOI: 10.1038/nrn3256.

[11] B. B. Biswal, M. Mennes, X. N. Zuo, *et al.*, "Toward discovery science of human brain function," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 10, pp. 4734–4739, 2010-03, ISSN: 00278424. DOI: 10.1073/pnas.0911855107.

[12] K. Stankevičiūte, T. Azevedo, A. Campbell, R. Bethlehem, and P. Liò, *Population graph GNNs for brain age prediction*, 2020-06. DOI: 10.1101/2020.06.26.172171.

[13] S. Arslan, S. I. Ktena, B. Glocker, and D. Rueckert, "Graph Saliency Maps through Spectral Convolutional Networks: Application to Sex Classification with Brain Connectivity," 2018-06. arXiv: 1806.01764. [Online]. Available: http://arxiv.org/abs/1806.01764.

[14] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *CoRR*, vol. abs/1609.02907, 2016. arXiv: 1609.02907. [Online]. Available: http://arxiv.org/abs/1609.02907.

[15] T. N. Kipf and M. Welling, *Variational graph auto-encoders*, 2016. arXiv: 1611.07308 [stat.ML].

[16] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *CoRR*, vol. abs/1901.00596, 2019. arXiv: 1901.00596. [Online]. Available: http://arxiv.org/abs/1901.00596.

[17] L. Jansson and T. Sandström, *Graph convolutional neural networks for brain connectivity analysis*, Master Thesis in Complex Adaptive Systems, Chalmers University of Technology, 2020.

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," 2017-10. arXiv: 1710.10903. [Online]. Available: http://arxiv.org/abs/1710.10903.

[19] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How Powerful are Graph Neural Networks?," 2018-10. arXiv: 1810.00826. [Online]. Available: http://arxiv.org/abs/1810.00826.

[20] F. Biagini, G. Kauermann, and T. Meyer-Brandis, "Introduction," in *Network Science: An Aerial View*, F. Biagini, G. Kauermann, and T. Meyer-Brandis, Eds. Cham: Springer International Publishing, 2019, pp. 1–4, ISBN: 978-3-030-26814-5. DOI: 10.1007/978-3-030-26814-5_1. [Online]. Available: https://doi.org/10.1007/978-3-030-26814-5_1.

[21] O. Nagar, S. Frydman, O. Hochman, and Y. Louzoun, "Quadratic GCN for graph classification," *CoRR*, vol. abs/2104.06750, 2021. arXiv: 2104.06750. [Online]. Available: https://arxiv.org/abs/2104.06750.

[22] K. Madhawa and T. Murata, "Active learning for node classification: An evaluation," *Entropy*, vol. 22, no. 10, 2020, ISSN: 1099-4300. DOI: 10.3390/e22101164. [Online]. Available: https://www.mdpi.com/1099-4300/22/10/1164.

[23] W. D. Joyner and C. G. Melles, "Introduction: Graphs—basic definitions," in *Adventures in Graph Theory*. Cham: Springer International Publishing, 2017, pp. 1–39, ISBN: 978-3-319-68383-6. DOI: 10.1007/978-3-319-68383-6_1. [Online]. Available: https://doi.org/10.1007/978-3-319-68383-6_1.

[24] E. Cozzo, G. F. de Arruda, F. A. Rodrigues, and Y. Moreno, "Multiplex networks: Basic definition and formalism," in *Multiplex Networks: Basic Formalism and Structural Properties*. Cham: Springer International Publishing, 2018, pp. 7–20, ISBN: 978-3-319-92255-3. DOI: 10.1007/978-3-319-92255-3_2. [Online]. Available: https://doi.org/10.1007/978-3-319-92255-3_2.

[25] T. Funke, M. Khosla, and A. Anand, *Hard masking for explaining graph neural networks*, Under double blind review as a conference paper at ICLR 2021, 2021. [Online]. Available: https://openreview.net/forum?id=uDN8pRAdsoC.

[26] B. H. Kim and J. C. Ye, "Understanding Graph Isomorphism Network for rs-fMRI Functional Connectivity Analysis," *Frontiers in Neuroscience*, vol. 14, 2020-06, ISSN: 1662453X. DOI: 10.3389/fnins.2020.00630.

[27] UKBiobank. (2021). "Uk biobank brain imaging – online resources," [Online]. Available: https://www.fmrib.ox.ac.uk/ukbiobank/. Accessed 2021-04-14.

[28] UKBiobank. (2021). "Uk biobank – enabling sciencitif discoveries that improve human health," [Online]. Available: https://www.ukbiobank.ac.uk/. Retrieved 2021-04-28.

[29] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[30] F. Chollet *et al.*, *Keras*, https://github.com/fchollet/keras, 2015.

[31] Martín Abadi, Ashish Agarwal, Paul Barham, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/.

[32] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG].

[33] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, "Accurate brain age prediction with lightweight deep neural networks," *Medical Image Analysis*, vol. 68, 2021-02, ISSN: 13618423. DOI: 10.1016/j.media.2020.101871.

[34] H. Jiang, P. Cao, M. Y. Xu, J. Yang, and O. Zaiane, "Hi-GCN: A hierarchical graph convolution network for graph embedding learning of brain network and brain disorders prediction," *Computers in Biology and Medicine*, vol. 127, 2020-12, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2020.104096.

[35] B. M. Mathisen, A. Aamodt, K. Bach, and H. Langseth, "Learning similarity measures from data," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 129–143, 2020-06, ISSN: 21926360. DOI: 10.1007/s13748-019-00201-2. arXiv: 2001.05312.

[36] C. Zhang, N. D. Cahill, M. R. Arbabshirani, T. White, S. A. Baum, and A. M. Michael, "Sex and Age Effects of Functional Connectivity in Early Adulthood," *Brain Connectivity*, vol. 6, no. 9, pp. 700–713, 2016-11, ISSN: 21580022. DOI: 10.1089/brain.2016.0429. [Online]. Available: /pmc/articles/PMC5105352/%20/pmc/articles/PMC5105352/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5105352/.

[37] J. Song, R. M. Birn, M. Boly, *et al.*, "Age-related reorganizational changes in modularity and functional connectivity of human brain networks," *Brain connectivity*, vol. 4, no. 9, pp. 662–676, 2014-11, ISSN: 21580022. DOI: 10.1089/brain.2014.0286.

[38] D. Tomasi and N. D. Volkow, "Aging and functional brain networks," *Molecular Psychiatry*, vol. 17, no. 5, pp. 549–558, 2012, ISSN: 14765578. DOI: 10.1038/mp.2011.81.

[39] T. B. Meier, A. S. Desphande, S. Vergun, *et al.*, "Support vector machine classification and characterization of age-related reorganization of functional brain networks," *NeuroImage*, vol. 60, no. 1, pp. 601–613, 2012-03, ISSN: 10538119. DOI: 10.1016/j.neuroimage.2011.12.052.

[40] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in Graph Neural Networks: A Taxonomic Survey," 2020-12. arXiv: 2012.15445. [Online]. Available: http://arxiv.org/abs/2012.15445.

[41] P. E. Pope, S. Kolouri, M. Rostami, C. E. Martin, and H. Hoffmann, "Explainability methods for graph convolutional neural networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, 2019, pp. 10 764–10 773, ISBN: 9781728132938. DOI: 10.1109/CVPR.2019.01103.

A. Model architectures

This chapter goes into more detail on the specific architectures for the models presented in Chapter 3. Specifically, the output shape, number of parameters and the connections for each layer in all models is presented in the form of Tables A.1, A.2, A.3 and A.4. When numbers are presented for either output shape or number of parameters they refer to age/sex.

Table A.1: Baseline

Layer	Output shape	Params	Connected to
Input layer A	(210)		
Dense layer	(1/2)	211/422	Input layer A
Total params		211/422	

Table A.2: GCN

Layer	Output shape	Params	Connected to
Input layer popgraph	(42, 42)		
Input layer X	(42, 42)		
GCN layer 1	(42,10)	430	Input layer popgraph Input layer X
GCN layer 2	(42,10)	110	GCN layer 1
GCN layer 3	(42,10)	110	GCN layer 2
Concatenate	(42, 30)	0	GCN layer 1 GCN layer 2 GCN layer 3
Dense layer	(1/2)	1261/2522	Concatenate
Total params		1911/3172	

Table A.3: Poptoy

Layer	Output shape	Params	Connected to
Input layer popgraph	(100, 100)		
Input layer X	(100, 100)		
GCN layer 1	(100,10)	20	Input layer popgraph Input layer X
GCN layer 2	(100,10)	110	GCN layer 1
GCN layer 3	(100,10)	110	GCN layer 2
GCN layer 4	(100,10)	110	GCN layer 3
GCN layer 5	(100,10)	110	GCN layer 4
Concatenate	(100, 50)	0	GCN layer 1 GCN layer 2 GCN layer 3 GCN layer 4 GCN layer 5
Dense layer 1	(100, 32)	1632	Concatenate
Dense layer 2	(100, 16)	528	Dense layer 1
Dense layer 3	(100, 1/2)	17/34	Dense layer 2
Total params		3627/3644	

Table A.4: Popencoder

Layer	Output shape	Params	Connected to
Input layer A	(100, 42,42)		
Input layer X	(100, 42,42)		
Input layer popgraph	(100,100)		
Encoder GCN layer 1	(100,42,10)	430	Input layer A Input layer X
Encoder GCN layer 2	(100,42,10)	110	Encoder GCN layer 1
Concatenate 1	(100, 42, 20)	0	Encoder GCN layer 1 Encoder GCN layer 2
Dense encoder	(100,1/2)	841/1682	Concatenate 1
GCN layer 1	(100,10)	1010	Input layer popgraph Dense encoder
GCN layer 2	(100,10)	110	GCN layer 1
GCN layer 3	(100,10)	110	GCN layer 2
GCN layer 4	(100,10)	110	GCN layer 3
GCN layer 5	(100,10)	110	GCN layer 4
Concatenate 2	(100, 51/52)	0	GCN layer 1 GCN layer 2 GCN layer 3 GCN layer 4 GCN layer 5
Dense layer 1	(100, 32)	1664/1696	Concatenate 2
Dense layer 2	(100, 16)	528	Dense layer 1
Dense layer 3	(100, 1/2)	17/34	Dense layer 2
Total params		4050/4950	

B. Model variability in Zorro

In this chapter, a short investigation into the model uncertainty in Zorro for GCN for sex prediction will be presented. In Figure B.1, standard deviation in the importance score for each node from two different versions of Zorro are presented. $\sigma_{\text{subjects}}(\mathcal{I})$ refers to the standard deviation when Zorro was run for a single GCN model, where the Zorro results were grouped into 23 different groups of 200 subjects each, with the uncertainty evaluated over the mean importance for each of the 23 different groups. This corresponds to the uncertainty presented for GCN for sex prediction in Section 4.2.1. $\sigma_{\text{models}}(\mathcal{I})$ refers to running Zorro for a single group of 200 subjects, but repeated for ten different model initialisations of GCN. Comparing the results, we observe that the uncertainties are of comparable size, but that $\sigma_{\text{subjects}}(\mathcal{I})$ is generally slightly higher except for some nodes. For this reason, we draw the conclusion that the uncertainties in Zorro with regards to varying model, $\sigma_{\text{models}}(\mathcal{I})$, are not negligible but are smaller than the uncertainty when varying subjects, $\sigma_{\text{subjects}}(\mathcal{I})$. Thus, the more interesting uncertainty is the one regarding subject uncertainty, but varying model uncertainty would be interesting to investigate in a possible future work where time is not an issue.

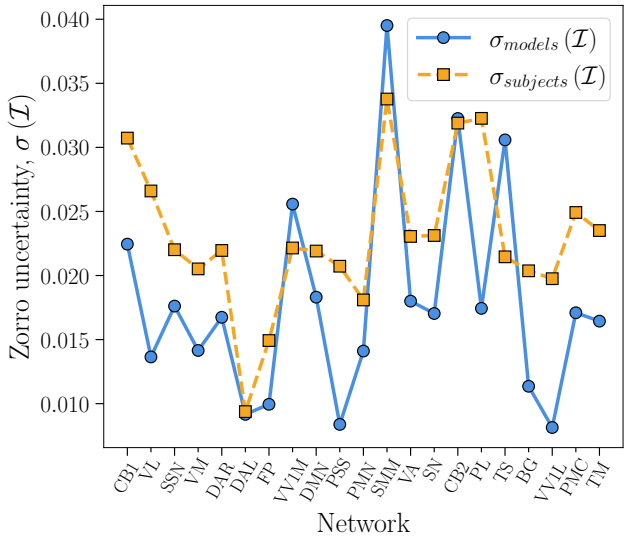


Figure B.1: Zorro uncertainty in the case of varying subjects for a single model, $\sigma_{\text{subjects}}(\mathcal{I})$, and in the case of a single group of subjects but for ten different model initialisation, $\sigma_{\text{models}}(\mathcal{I})$. Observe that $\sigma_{\text{subjects}}(\mathcal{I}) > \sigma_{\text{models}}(\mathcal{I})$ for most nodes.

DEPARTMENT OF PHYSICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY