Master Thesis

Exploring Brain Connectivity in Ageing Using Explainable Deep Learning

Name:	Tamar van Asch
StudentID:	0906621
Faculty:	Biomedical Engineering
Master:	Medical Engineering
Group:	Medical Image Analysis
Course Code:	8ZM205 & 8ZM91
Organization:	Department of Radiology, Maastricht University Medical Centre
Duration:	September 2020 - September 2021 (60 ECTS)

Supervision:

Prof.dr. Marcel Breeuwer	Department of Biomedical Engineering, Eindhoven University of Technology
Dr. Jacobus Jansen	Department of Radiology, Maastricht University Medical Centre
Nathan de Jong	Department of Radiology, Maastricht University Medical Centre

Committee:

Prof.dr. Marcel Breeuwer	Department of Biomedical Engineering, Eindhoven University of Technology
Dr. Jacobus Jansen	Department of Radiology, Maastricht University Medical Centre
Dr. Mitko Veta	Department of Biomedical Engineering, Eindhoven University of Technology
Dr. Bart Markvoort	Department of Biomedical Engineering, Eindhoven University of Technology

This work was performed in compliance with the TU/e scientific code of conduct

Exploring Brain Connectivity in Ageing Using Explainable Deep Learning

T.G. van Asch (0906621)

Abstract—Population ageing increases the burden of age-related functional decline and disease. Consequently, there is a growing need for the understanding of the process of ageing and the ability to predict who is at risk of neurodegenerative diseases and mortality. A biomarker for the underlying biological process of ageing in the brain is brain age. Using non-invasive neuroimaging data in a deep learning framework, brain age can be accurately predicted. The goal of this research is to train a convolutional neural network (CNN) on structural brain connectivity data to predict age. Additionally, brain connectivity in ageing is explored using two explainable deep learning methods: sensitivity analysis and Integrated Gradients. Diffusion-weighted magnetic resonance imaging (dMRI) is used to map the structural connectivity of the brain. The dataset comprises 3494 subjects from The Maastricht Study, a cohort study of individuals aged between 40 and 77 years. Brain age prediction on the test set resulted in a Pearson's correlation coefficient of 0.700 and a mean absolute error (MAE) of 5.11 years. The analysis of the input data indicates that multiple brain regions have a large contribution to the prediction of age. These regions include the superior and medial frontal lobe, the basal ganglia and the precuneus. Further research is necessary to investigate the relation between these regions and the ageing process, as well as the consistency of the found contributions.

Index Terms—Brain age, deep learning, diffusion-weighted magnetic resonance imaging (dMRI), Integrated Gradients, sensitivity analysis, structural connectivity

1 Introduction

Inevitably, we undergo the process of time-dependent functional decline, also known as ageing [1]. Ageing has attracted curiosity and driven researchers throughout history to get a better understanding of this complex process. Even more so, how to delay, prevent or reverse its negative effects. Nevertheless, amongst today's society, the proportion of older people is rapidly increasing. It is expected that by 2050 approximately 22% of the world population will be 60 years or older [2]. Population ageing increases the burden of age-related functional decline and disease [3]. Consequently, there is a growing need for the understanding of the process of ageing and the ability to predict who is at risk and which treatments are applicable. For the understanding of the reader, current research on ageing in the brain will be discussed in the next section along with a brief recap on the anatomy of the brain.

1.1 Neuroanatomy of the (ageing) brain

Three main structures can be identified in the human brain. Firstly, the brain stem, which is responsible for automatic behaviours that are necessary for survival and is the first to develop. Secondly, the limbic system, which is our emotional brain and is extensively connected to lower and higher brain regions. Thirdly, the cerebral hemispheres, which are the most conscious parts of the brain and contain the primary motor cortex, primary sensory cortex, and the association cortices or higher cognitive functions of the brain. The cerebral hemispheres form the superior part of the brain and consist of ridges of tissue are called gyri, which are separated by shallow grooves called sulci (Fig. A.1a, Appendix A). The gyri and sulci form landmarks that are used to further subdivide the hemispheres into four lobes: frontal, parietal, temporal and occipital (Fig. A.1a, Appendix A) [4].

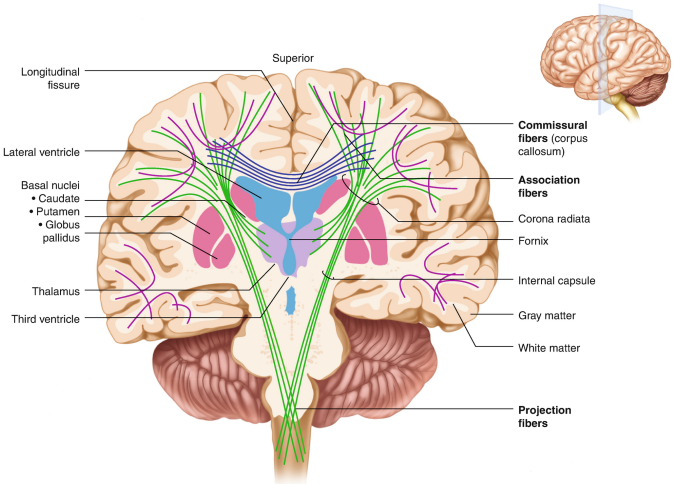


Fig. 1: Schematic view of the three different types of white matter tracts: association fibers (purple) that run between regions in the same hemisphere, commissural fibers (blue) that connect both hemispheres and run mainly through the corpus callosum and projection fibers (green), which pass between the thalamus and basal nuclei. Image is originally from [4].

The functional elements in our nerves system are the neurons which are cells that receive and transmit electrical signals throughout our body. Within the brain a distinction can be made between white and grey matter, where grey matter is a collection of neuronal bodies and white matter contains the tracts over which the signals are transferred [5]. White matter is responsible for communication between cerebral areas and can roughly be divided in three different types of fibers: the association fibers connect different parts of the same hemisphere, commissural fibers allow the two hemispheres to communicate with each other and lastly the

projection fibers either run down to the rest of the body or enter the cortex from lower brain regions, see Fig. 1 [4].

In general, ageing of the brain is suggested to reverse the sequence of development [6]. This is referred to as the last-in-first-out hypothesis [7], [8]. White matter tracts develop through myelination, which is a process that increases the conduction velocity of electrical signals, improving brain connectivity. The majority of myelination occurs in the first 2 years of life, but the process continues until your twenties [9]. Generally, projection fibers emerge early while association fibers are relatively late to mature [10]. The hypothesis suggests that tracts that are late to mature, and in particular late-myelinating fibers, are more vulnerable to injury and decline in later life [6], [11]. Subsequently, researchers report that association tracts deteriorate more than commissural tracts [12]. Brain structures that myelinate last include the association cortices and neostriatium. The neostriatum is part of the basal ganglia (Fig. A.1c, Appendix A), and collects information from the four lobes of the cerebral hemispheres [13]. Accordingly, it has been shown that the association cortices and neostriatium are indeed more vulnerable to the effects of ageing [8].

1.2 Predicting brain age

Genetic and environmental factors cause a large variation in the way we age and the timing with which we develop age-related diseases [14]. Neurodegenerative diseases such as Alzheimer's disease accelerate ageing of the brain [15]. Factors that have a positive influence on brain ageing are physical exercise, years of education and practicing meditation [16], [17], [18]. A biomarker for the underlying biological process of ageing in the brain has previously been introduced by Cole et al. and can be used to identify individuals deviating from a healthy brain age trajectory. This biomarker is called brain age and is derived from neuroimaging data [16]. Using non-invasive neuroimaging data in a machine or deep learning framework, brain age can be predicted with an accuracy up to 2.5 years [19]. Research on brain age has shown that 'older brains' in adults indicate an increased risk of neurodegenerative diseases and mortality [14]. On the other hand, investigation of biological younger brains can support us in the understanding of healthy ageing and the actions we have to take to preserve, prevent or delay age-related effects [20]. Consequently, the development of accurate brain age prediction models can lead to a diagnostic tool for neurodegenerative diseases and the understanding of the effects of health characteristics and life experiences on brain ageing [21].

1.3 Explainable predictions

Besides accurately predicting brain age, it is desirable to investigate the subjects that have a large difference in their chronological and brain age, since it can clarify which factors contribute to accelerated or delayed ageing. Currently, deep learning networks outperform regular regression and machine learning models in brain age prediction [19]. Deep learning networks take an input or a set of input features and extract higher nonlinear features from this data. Due to multiple layers of artificial neurons, deep learning networks are able to automatically recognize complex patterns in imaging data [22], [23]. However, one of the main issues in deep learning is its black-box aspect. It is often unknown what information drives the decisions of the network. This has opened a new research field: explainable deep learning (XAI). Here, a distinction can be made between explainability of the information flow inside the network and explainability of the input data. In this research the focus will be on the latter. From a clinical point of view, it is more interesting to explain what information drives the decision of the network. Also, examination of the input data can guide improvements in the pre-processing of the data.

XAI has the potential to explain the predicted brain age. For example, Levakov et al. predicted brain age with an average accuracy of three years and showed that cavities containing cerebrospinal fluid had the highest contribution for brain age prediction [24]. However, comparing multiple studies on brain age prediction shows that the results are inconsistent [19], [25], [26]. An explanation for this inconsistency is the complexity of neural networks, meaning that assigning a contribution to each input feature is not evident. Furthermore, the lack of a ground truth makes it difficult to assess XAI methods and their outcomes quantitatively. It is not obvious to determine if artifacts are caused by a misbehaving network or a misbehaving attribution method [27]. This has led to the development of several XAI methods [28], [29], [30], [31], [32]. However, there is no golden standard and further exploration of the XAI results is necessary before it can be implemented in a clinical setting.

Additionally, most studies on XAI in brain age prediction use regular T1-weighted magnetic resonance imaging (MRI) [19], [24], [25], [26]. However, subtle age-related changes might not be captured by regular MRI sequences [33]. Studies have shown that white matter microstructure integrity degrades due to the effects of ageing [34], [35]. An advanced MRI technique that is used to characterize white matter changes in neurodegeneration is diffusion-weighted MRI (dMRI) [36]. dMRI measures the diffusion of water molecules to determine fiber tract orientation and tissue anisotropy. The pathways of the white matter fiber tracts can be calculated with tractography, a three-dimensional (3D) modelling technique [37], [38]. Accordingly, this allows mapping of the structural connectivity of the brain. Prediction of brain age on a small dMRI dataset has recently been done by Chen et al., which resulted in a prediction error of approximately five years [33].

1.4 Research purpose

The goal of this research is to explore structural connectivity in ageing by implementation of XAI methods. As described above, brain age prediction models are widely investigated, however, they lack information on the microstructure of white matter tracts. Within this study, we have the unique opportunity to work with more than 3000 subjects that underwent an extensive MRI scanning protocol, including dMRI. This dataset makes it possible to train a convolutional neural network (CNN) to predict brain age.

This study starts with testing the feasibility of brain age prediction using a deep learning network or model that is trained on dMRI derived connectivity data. Data acquisition and pre-processing was previously done to this study. For

reproducibility purposes, a summary of the dataset and the protocol for acquisition and pre-processing can be found in section 2.1 and 2.2. Experiments are conducted to optimize the CNN that is used for the prediction of brain age. The results are documented in Appendices B and C. The parameter settings that led to the best performance are used for the final predictions and are described in section 2.3 and 2.4. Afterwards, two XAI methods are implemented and used to compute the relative contribution of each input feature, which is described in section 2.5. The input features that are found to be most predictive of age are compared with literature and compared with the change in connectivity with increasing age to verify the results of both XAI methods. Lastly, the consistency of the results is tested. An overview of the technical terms that are used throughout this report can be found in Appendix A.

2 METHODOLOGY

2.1 Dataset

The data used in this research is from The Maastricht Study, a population-based cohort study oversampled on type 2 diabetes mellitus in midlife individuals aged between 40 and 77 years. The sampling method and protocol have been described previously [39]. The Maastricht Study started late in 2010, and it now includes over 10,000 individuals, of whom more than half underwent an extensive MRI scanning protocol.

The MRI data is acquired on a 3T clinical magnetic resonance scanner (MAGNETOM Prismafit, Siemens Healthineers GmbH, Munich, Germany) located at the scanning facility Scannexus in Maastricht, The Netherlands. A head/neck coil is used with 64-elements for parallel imaging. The MRI protocol includes acquisition of multiple sequences. Acquisition of dMRI is done using a diffusion sensitized echo-planar imaging (EPI) sequence (TR/TE $6100/57ms$, 65 slices, 100×100 matrix size, 64 diffusion sensitizing gradient directions ($b = 1200\,s/mm^2$), $2.0\,mm$ cubic reconstructed voxel size) with additionally three minimally-diffusion-weighted images ($b = 0\ s/mm^2$) [40].

Within this research, a dataset that comprises 3494 subjects from The Maastricht Study is used. Besides dMRI data, age, gender, diabetes status and education level of all subjects are available. Additionally, all subjects have previously been evaluated on the presence of symptoms of cerebral small vessel disease (CSVD). Each subject is assigned a CSVD score based on white matter hyperintensities volume, presence of microbleeds and lacunar infarcts [41]. The assigned score is in range $[0, 3]$, where a score of zero indicates the absence of all three, and a score of three indicates the presence of all three. Subjects with a score of 2 or 3 are excluded from this study.

The subjects are divided over a training (85%) and independent test (15%) cohort. From the training set 20% is used for validation. Age is non-uniformly distributed with a mean age of 59 years and a standard deviation of 8.6 years. Half of the participants are female. Table I shows the distribution of the participants over the training and test cohort.

TABLE I: The distribution of data within the training and test cohort. STD = standard deviation, T2D = Type 2 Diabetes.

	Training Cohort	Test Cohort
Number of subjects	2984	510
Mean age (years)	60.4	60.4
STD age (years)	8.68	8.96
Percentage female	50.2	48.9
Percentage T2D	19.8	22.5

2.2 Structural connectivity

2.2.1 Pre-processing

Within this research, dMRI data is used to predict brain age. Raw data obtained from the protocol, as described in the previous section, is pre-processed towards a compact 2D matrix that describes the structural connectivity. The pre-processing was already done by the protocol of De Jong et al. [40]. The raw dMRI data is anonymized and converted from DICOM to NIfTI format. Most pre-processing steps are done using ExploreDTI v4.8.6 (PROVIDI lab, Image Sciences Institute, Utrecht, The Netherlands) [42] and included eddy current and head motion correction [43], [44]. Additionally, constrained spherical deconvolution based deterministic whole-brain tractography was applied to acquire the white matter tracts [45].

2.2.2 From a 3D Tractogram to a 2D Brain Network

The cerebrum can be parcelled into 94 brain regions using the automated anatomical labelling (AAL) atlas [46]. Regions are connected via white matter tracts, which are mapped using dMRI. Each brain region from the 3D tractogram can be represented by a node in a 2D brain network, which are the dots in the left-hand side of Fig. 2. The white matter tracts that are reconstructed using dMRI, represent connections between a pair of nodes. The connection between two nodes can be represented by an edge, visualized as straight lines in Fig. 2. Using dMRI data, connection strength can be calculated. For each pair of brain regions that are connected through two or more tracts, the connection strength is determined by the tract volume (the number of voxels visited by a tract multiplied by voxel size) relative to intracranial volume [40]. Connection strength corresponds to the line thickness of the edges in Fig. 2. Additionally, each node can be characterized by its strength, which is defined as the sum of the connection strength of each connection to that node [47]. The node strength corresponds to the size of the dots in Fig. 2.

2.2.3 Graph Networks

The brain is a complex system of interconnected regions that can thus be described by a set of nodes and edges. This characterization of the brain as a network is derived from the mathematical study of networks: graph theory [47]. A graph network G consists of a non-empty finite set of nodes $\Omega(G)$ and a finite set of edges $E(G)$. An edge $\{i, j\}$ joins nodes i and j. From this, an adjacency matrix $\mathbf{A}$ can be constructed of which the ij-th entry represents the edge that joins nodes i and j [48].

The brain can be constructed as a graph network $G = (\mathbf{A}, \Omega)$ in which Ω is a set of nodes representing all brain regions of interest (equal for all subjects) and $\mathbf{A} \in R^{|\Omega|x|\Omega|}$

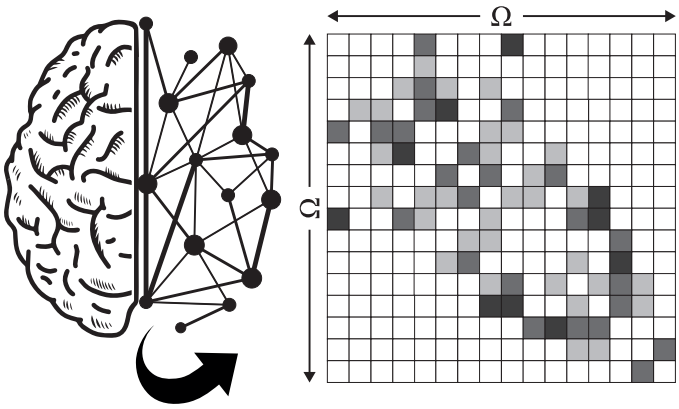


Fig. 2: Visualization of the construction of a brain network with the human brain as a graph network (left) and its corresponding weighted adjacency or connectivity matrix (right). For each region in the set of brain regions Ω, the connection strength with every other region in Ω is represented using different colour intensities. Here, self-connections are put to zero (represented by white elements in the image).

is a weighted adjacency matrix of edges that represent connection strength [49]. In this research, a weighted adjacency matrix is referred to as a connectivity matrix. Fig. 2 shows the construction of a virtual connectivity matrix. The intensity of the colour corresponds with the connection strength.

2.3 Network Architecture

This research proposes a CNN to predict brain age from structural connectivity data. The input for the network is a 94 x 94 connectivity matrix. Although connectivity matrices can easily be mistaken for an image, regular convolution would undermine the topology of the matrix. Surrounding elements in the matrix are not per definition in close proximity in the brain. Therefore, Kawahara et al., implemented a novel design of the first convolution layer of the network: a cross-shaped kernel that does not compromise the topology of the connectivity matrix [49].

The first layer is an edge-to-edge convolution filter containing the cross-shaped kernel that combines the edges that share nodes. This kernel sums all edges along row i and column j. The second convolution layer is an edge-to-node layer, which is a 1D convolution summing the i_{th} row, and therefore, an equivalent of node strength is calculated. The third convolution layer is again a 1D convolution layer that sums all the nodes and generates a score per graph per filter [49].

The architecture and hyperparameter settings are mainly based on the original BrainNetCNN architecture [49]. Later publications reduced the amount of layers and filters [50], [51]. Therefore, the chosen architecture contains three convolution layers and two fully connected layers, and the amount of filters has been decreased by half, see Fig. 3. Each layer uses the leaky rectified linear units (ReLU) as an activation function with a slope of 0.33 for all values < 0.

Additionally, experiments are designed to optimize the network's performance. Appendix B contains an overview of the training results on these different hyperparameter settings. Variations in batch size, loss-function, dropout and weight decay were tested.

2.4 Training Procedure

The network is implemented using Keras v2.4.2 [52] with TensorFlow v2.4.1 backend [53]. The network is trained using the mean squared error loss function and is asked to predict a score in range $[0, 1]$ in which 0 corresponds to an age of 40 years (the youngest subjects) and 1 corresponds to 77 years (the oldest subjects). An epoch is defined as an iteration over all batches and the network is trained for a maximum of 1000 epochs with a batch size of 45. Training is terminated early if the validation loss has not improved within the last 100 epochs. Training is done using the stochastic gradient descent optimizer with an initial learning rate of 0.01. An additional term in the loss function, the L2 regularization, forces the trainable parameters to stay small. The weight decay used is 0.0005.

The network is trained in a five fold cross validation, which means that it is independently trained five times. The training cohort is divided into five folds and for each training a different set is used as validation set. The performance of the network is evaluated using an independent test set. The predictions on this test set are obtained by averaging the prediction of all five networks from the cross-validation training, which is called the ensemble performance.

2.4.1 Data normalization

The data is normalized to optimize the training of the network by applying equation 1 to each subject's connectivity matrix $\mathbf{A}_{ij}$. The maximum value $max(\mathbf{A})$ is determined by the 99.9% percentile to prevent outliers from strongly influencing the resulting matrix. The connectivity matrix only contains positive values and therefore, the minimum value $min(\mathbf{A})$ equals zero for all matrices. Additionally, self-connections are removed and put to zero.

$$\mathbf{A}_{ij,scaled} = \frac{\mathbf{A}_{ij} - min(\mathbf{A})}{max(\mathbf{A}) - min(\mathbf{A})} \tag{1}$$

2.4.2 Data augmentation

Data augmentation is a well-known regularization technique in image classification and has a significant contribution in the optimization of neural networks. Data augmentations apply label-preserving transformations to the input data such as rotation, skewing and translation which effectively enlarges the training set. However, these transformations are not applicable to graph networks. This is mainly due to the lack of spatial locality in graphs, and unfortunately, alternative methods are under-explored [54], [55]. Most techniques for graph networks focus on augmenting graph structures by adding or removing edges [56].

In Appendix C, the results of different augmentation methods are reported. The best performing method with the lowest computational costs is the Keras Gaussian Noise layer, which is added to the network to reduce overfitting. The build-in Gaussian Noise layer in the Keras environment adds random noise from a normal distribution with mean $\mu = 0$ and a standard deviation σ that can be tuned with the other hyperparameters. To mimic data augmentation, this layer is applied on the input data. The network is less able to memorize the data as it slightly changes every time. This results in smaller network weights and a network that

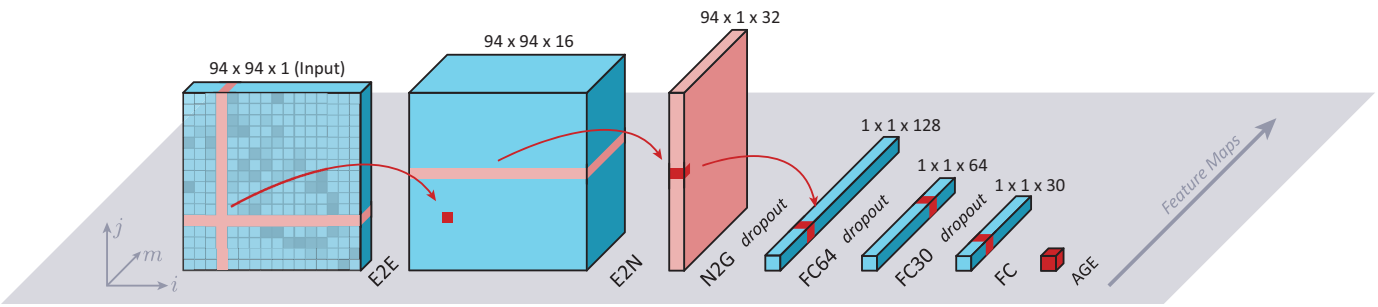


Fig. 3: Schematic representation of the network. Image is based on [49]. The first layer is the edge-to-edge (E2E) convolution layer that sums the edges of node i and j. The second layer is an edge-to-node (E2N), which consists of a 1D convolution filter, and the third convolution layer is a node-to-graph (N2G) layer, which again consists of a 1D convolution filter. The final layers of the network are fully connected (FC) layers. The final output of the network is the prediction of brain age. The convolution filters are visualized in pink and the amount of feature maps m is given in the image.

has a lower generalization error. Due to the exponential distribution of the data (Fig. C.1, Appendix C), a standard deviation of 0.01 is chosen (approximately 10% of most data points).

2.4.3 Evaluation Metrics

Similar to previously mentioned literature, the mean absolute error (MAE) and the Pearson's correlation coefficient (ρ) will be used to evaluate the performance of the network. An MAE score can be disproportionally low if the network predicts regularly nearer to the mean score of the training data. Since most subjects are aged close to the mean, the Pearson's correlation coefficient is used as an additional evaluation metric. The Pearson's correlation coefficient measures the linear dependence between predicted and chronological age and is less affected by the distribution of the input data [49].

2.5 Explainable Deep Learning

This research aims to get a better understanding of the contribution of each input feature to the output of the network. This is achieved by implementation of two XAI methods. The contribution of an input point to the output is called its attribution. The visualization of the attributions in the shape of the input matrix will be referred to as attribution maps (in image classification models this is often referred to as a heatmap).

2.5.1 Gradient based attribution methods

In this research, two attribution methods are implemented and compared: sensitivity analysis and Integrated Gradients. The methods can be applied post-model and therefore do not require any modifications to the architecture of the network. In section 4.2, the results will also be compared to the literature.

The gradient with respect to the input features is the fundamental piece of information in both methods. Starting with a linear regression model, the role of gradients in attribution methods can be clearly explained. Given input x_i, output y and w_i as the learned model weights, a linear model can be described by equation 2.

$$y = w_1x_1 + w_2x_2 + ... + w_nx_n \tag{2}$$

Calculating the partial derivative $\frac{\partial y}{\partial x_1}$ results in w_1 which suggests that the magnitude of w_i, i.e. the gradient of x_i, determines the attribution of the corresponding element x_i to y. This implies that for a large gradient, a small deviation still has a large impact on the output. Gradients close to zero indicate no contribution to the output. To apply the same principal on CNNs, the partial derivative of the target output with respect to the inputs w_i can be calculated. This results in an expression given by equation 3, which can be seen as the first-order Taylor decomposition of the network function [27], [57].

$$w_i = \frac{\partial y}{\partial x_i} \tag{3}$$

Sensitivity analysis was first described by Simonyan et al. and applied previously to BrainNetCNN [28], [49]. It is conceptual a simple method, with low computational costs [58]. However, the results are expected to be rather noisy and incomplete [59]. Due to the sparsity of the connectivity matrices, the gradients are multiplied element-wise with the corresponding input features to obtain an attribution R, as described by equation 4. This method was previously proposed to improve the sharpness of the gradients and resulted in less noisy attribution maps [31].

$$R_i = \frac{\partial y}{\partial x_i} \cdot x_i \tag{4}$$

Similar to the sensitivity analysis, Integrated Gradients computes the partial derivative of the output with respect to each input feature. Instead of a single derivative per input feature, a series of inputs and corresponding gradients is generated by linear interpolation between a baseline $\bar{x}$ and the original input x. The baseline is chosen by the user and is required to be an 'information-less' matrix that turns off all the features. Usually a zero matrix is used. The mathematical description of the calculation of Integrated Gradients is given by equation 5, in which α is the number of interpolation steps. The total attribution is computed by taking the integral over the series of derivatives which is then multiplied by the difference between the input feature and its baseline. The method was first introduced by Sundararajan et al. [32] and has shown to produce clear attribution maps [60].

$$R_i(x) = (x_i - \bar{x}_i) \times \int_{\alpha=0}^{1} \frac{\partial y(\bar{x} + \alpha \times (x - \bar{x}))}{\partial x_i} \, d\alpha \tag{5}$$

The advantage of Integrated Gradients can be explained using Fig. 4. Due to the activation functions in CNNs, there is a chance that a change in the input is not detected, since the gradient of the activation function is zero at that point (for example the ReLU function for all input values lower than zero). This is referred to as the saturation effect. Intuitively, the saturation of the network leads to flat functions, making us believe that the gradient is zero and the feature is not important. By tuning down the intensity of the input, these changes can be detected. In the current network, Leaky ReLU functions are used, which do not force the gradients towards zero. Instead, they allow a small slope for input values smaller than zero. Hence, the Leaky ReLU activation causes damping of the gradient rather than complete saturation.

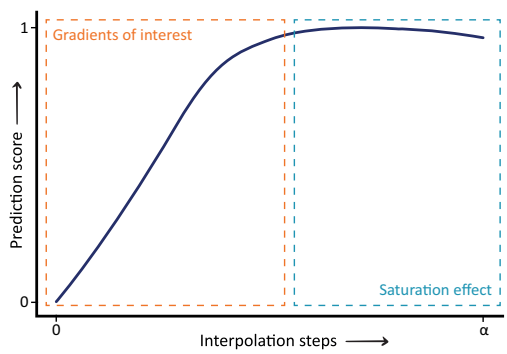


Fig. 4: The prediction score is expected to increase from 0 to 1 over the interpolation steps α. Left, the attributions of interest result in a steep increase of the prediction score. Right, a plateau represents the saturation effect caused by the activation functions.

Integrated Gradients was designed with the requirement that for $x_i \neq \bar{x}_i$ and $y(x) \neq y(\bar{x})$ a nonzero attribution is assigned. This is not the case in the sensitivity analysis, because of the saturation effect. Secondly, the method is invariant to the implementation, meaning that for two models that are functionally equivalent (same output given any input), their feature attribution will also be equivalent, which is desirable when investigating the input data rather than the model's behaviour [32].

2.5.2 Statistical Analysis

To evaluate the consistency of the attributions, the coefficient of variation (CV) is calculated for each edge, according to equation 6. The CV equals the ratio between the standard deviation σ and mean μ. The CV is used to express the inter-subject and inter-model variation per edge. The connections that are found to be predictive of age are expected to have a significant attribution in each subject and each model. Hence, a low variation is expected to be found for these connections.

$$CV = \frac{\sigma}{\mu} \tag{6}$$

2.5.3 Experiments

Using both the sensitivity analysis and Integrated Gradients as described above, an attribution map is computed for all subjects in the training cohort. The gradients are extracted using Tensorflow built-in function *GradientTape* for all subjects, for all five models. For the sensitivity analysis, the gradients are multiplied by the input and averaged over all subjects, per model. Then, the average attribution per input feature is calculated by taking the average over all five models.

The Integrated Gradients method is also applied on all subjects in the training cohort and an average attribution per input feature is calculated by averaging over all subjects and all five models, using a zero matrix baseline. Additionally, to mimic the increase in score over the interpolation steps from 0 to 1, an alternative baseline is chosen. This is a single subject that corresponds to the lowest prediction score. The Integrated Gradients method is then applied to elderly subjects aged $70 - 77$ (15% of the training set).

Despite the symmetry in the input data, the attribution maps are not symmetric around the diagonal. Therefore, the upper and lower triangle are summed to obtain the total attribution per edge. The attribution maps are normalized to range between $[-1, 1]$, using the 99.9 percentile maximum from the absolute values. A threshold is applied, which means that all attributions in the range $[-0.6, 0.6]$ are put to zero and all attributions larger than 1 or smaller than -1 are clipped to 1 or -1, respectively. The remaining attributions that remain will be referred to as the attributions of interest. BrainNet Viewer, a MATLAB (The MathWorks Inc., Natick, MA, US) based application, is used to visualize the attribution maps [61].

The input features that are found to be most predictive of age are compared with the average connection strength of the corresponding connections per age category. The goal is to verify if there is indeed a change with increasing age in the structural connectivity for these connections. Lastly, the inter-subject variation of the attributions is calculated per edge. The mean attribution and corresponding standard deviation over all subjects are used to calculate the CV. Additionally, the mean attribution over all subjects is calculated per model per edge. The variation in the mean attributions is the inter-model variation. Again, the CV is calculated to quantify this variation.

3 Results

Cross validation led to an average mean absolute error (MAE) of 5.41 ± 0.12 years and an average Pearson's correlation coefficient (ρ) of 0.661 ± 0.018. The ensemble performance of all five models improved to MAE $= 5.11$ years and $\rho = 0.700$. The loss curves of the training process can be found in Fig. D.1, Appendix D.

The age prediction results are shown in Fig. 5. A bias can be observed where lower aged subjects have a higher predicted age and vice versa. This bias is confirmed in the boxplot in Fig. 6. The outermost age groups show a larger prediction error than the age groups that are centered around the mean age of the dataset. From Fig. 6, it can also be observed that the model predicts most accurate for ages in range of 55 to 70 years.

3.1 Explainable Method

As described in section 2.5, three different attribution maps were computed. Fig. 7a shows the results of the sensitivity

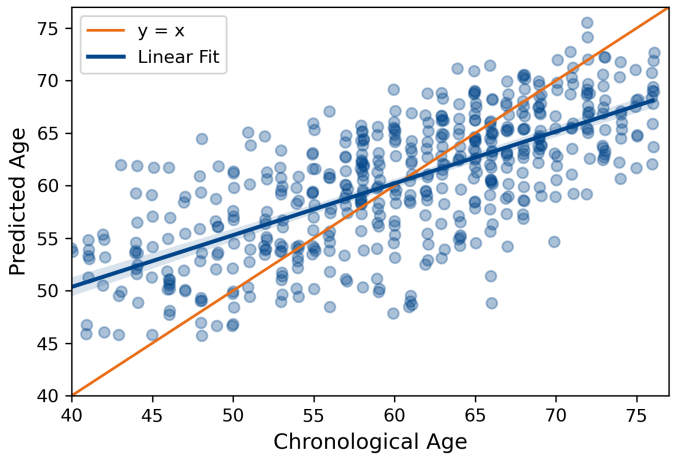


Fig. 5: A scatterplot of the ensemble predictions on the test set. The orange line shows the $y = x$ line in comparison to the blue linear fit through the predictions. A bias can be observed where lower aged subjects have a higher predicted age and vice versa.

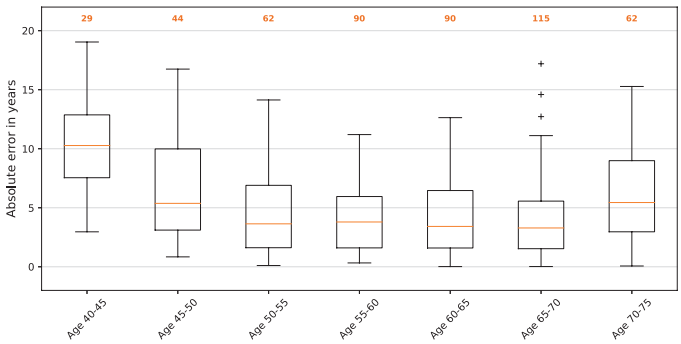


Fig. 6: The boxplot shows the absolute error distribution per age group of 5 years. The orange line in the box represents the median, the box defines the interquartile range (25% - 75%) and the whiskers are at the 5% confidence interval. The orange numbers in the top of the figure represent the amount of subjects in the corresponding age group.

analysis, Fig. 7b shows the results of the Integrated Gradients with a zero matrix baseline and Fig. 7c shows the results of the Integrated Gradients applied to the elderly, with the youngest predicted subject as baseline. All three attribution maps are shown from an axial (left) and coronal (right) point of view. The blue edges represent connections that have a negative influence on the prediction score. Whereas the red edges represent connections that positively influence the prediction score. A simplified interpretation is that the blue edges represent connections that decrease in strength over age, whereas the red edges represent connections that are expected to increase in strength over age. The thickness of the edges corresponds to the absolute value of the attribution. The node size corresponds to the node strength, which is calculated by taking the sum over the attributions per node.

The attribution maps in Fig. 7a and 7b greatly overlap. There are four strong blue edges. Interhemispheric between the precuneus (PCUN), the edge between the right precuneus and the left middle cingulate gyrus (MCG) and the edges between the dorsolateral superior frontal gyrus (SFGdor), insula (INS) and caudate (CAU) in the right hemisphere. The latter is stronger in Fig. 7a than 7b. The strongest red edge is between the PCUN and CMG in the left hemisphere. Another strong red edge can be found between the parahippocampal gyrus (PHG) and the fusiform gyrus (FFG), also in the left hemisphere.

Fig. 7c shows more interhemispheric edges. The CAU in both hemispheres has multiple strong edges. Also, most attributions are positive. Similar to the previous attribution maps, the SFGdor, PUT and pallidum (PAL) are present in both hemispheres.

To evaluate the connections that are indicated as contributing the most to the output, the average connection strength per age category can be visualized. The resulting graph for the attributions of the sensitivity analysis and Integrated Gradients with a zero baseline is shown in Fig. 8a. An additional threshold of 0.7 is applied on the attributions to exclude the attributions that do not overlap between both maps. For most negative attributions, it can be observed that the corresponding connections indeed decrease in connection strength with increasing age. However, this is not the case for the connection between the SFGdor and the INS. This connections appears to remain constant in connection strength with increasing age. The connections between the PUT and PAL correspond to positive attributions, and it can be seen that these connections are relatively strong and increase in strength with increasing age. Looking at Fig. 8b, the connection between the left and right CAU stands out. There is a rapid decrease in connectivity strength with increasing age. Only the connections between the PUT and PAL and the left and right thalamus increase with increasing age, the rest seems to decrease.

3.1.1 *Coefficient of Variation*

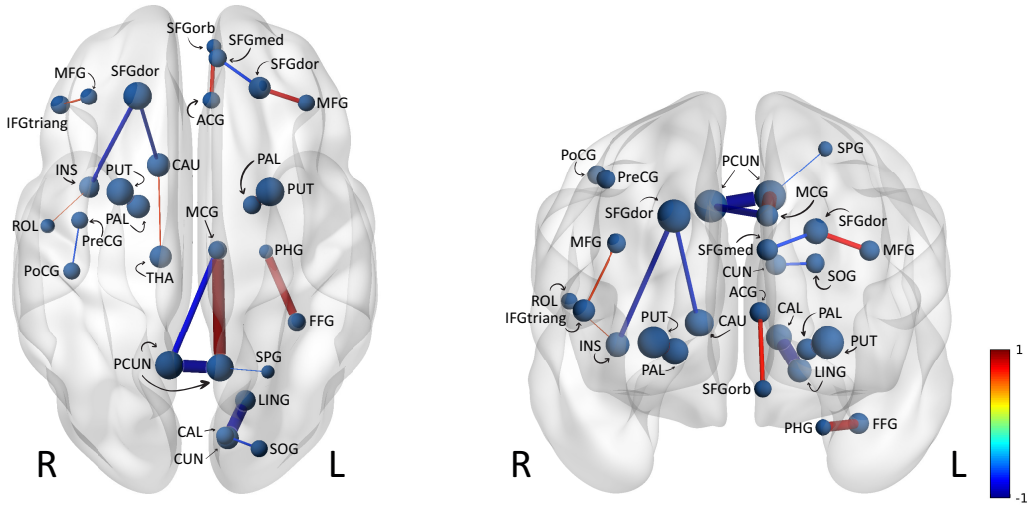
The inter-subject and inter-model variation of the attributions is calculated using the CV. For each edge, the CV is calculated. In this section, the median of the absolute values of the CV of all edges is reported. An overview of the results can be found in Table II. From this table, the Integrated Gradients using a zero matrix baseline appears to be the most consistent method with in general the lowest median CV. The CV per attribution of interest is summarized in Table D.I in Appendix D.

4 Discussion

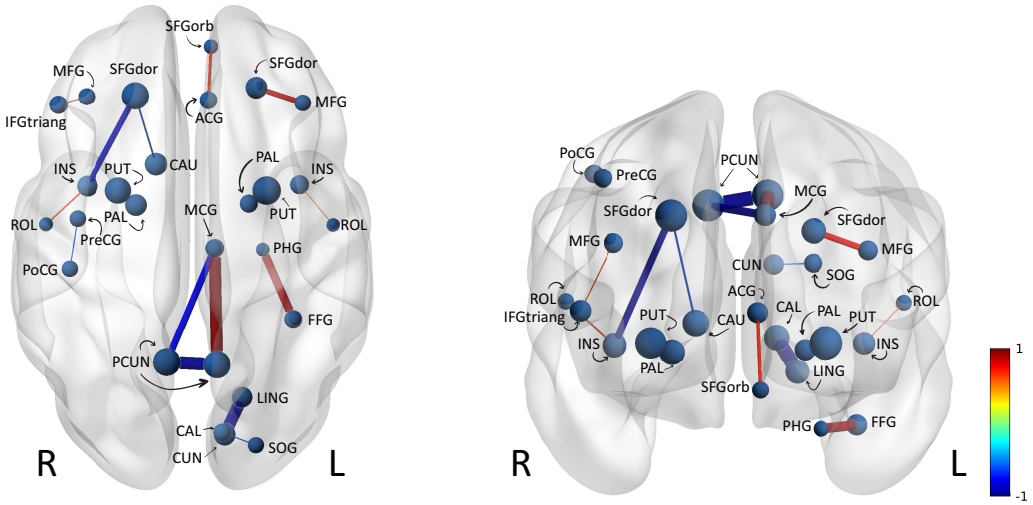
4.1 Network Performance

Within this research, a method was proposed to predict brain age based on structural connectivity data using a CNN. Additionally, two XAI methods have been implemented to examine the contribution of each input feature to the prediction score of the network. First, the performance of the network is discussed. The outcome of the XAI methods is discussed and compared to literature in section 4.2.

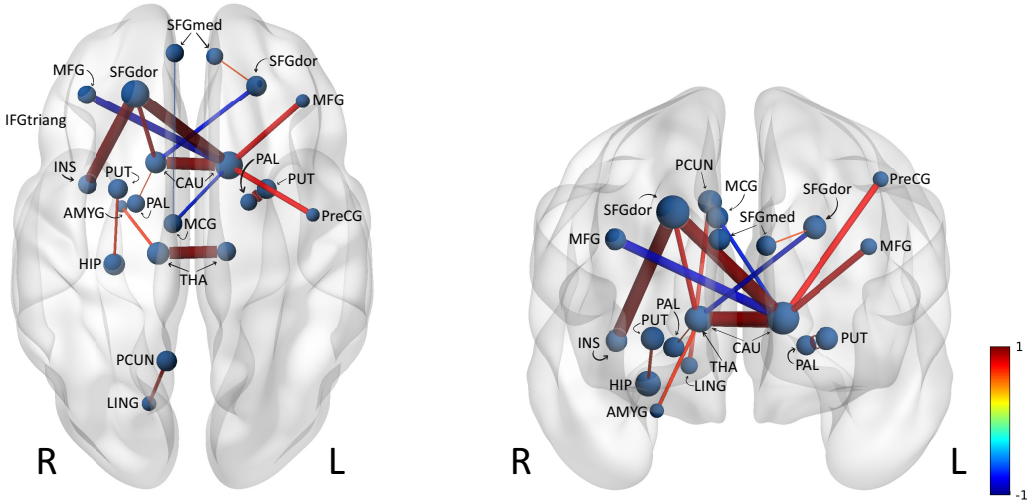
The ensemble performance of the proposed network is described by an MAE of 5.11 years and a Pearson's correlation coefficient of 0.700. The MAE is comparable with [33], which is, to our knowledge, the only to report brain age prediction using deep learning on dMRI derived features. However, the best reported MAE in state of the art models on brain age prediction on T1 weighted MRI is 2.15 years, and most studies report a Pearson's correlation coefficient $\rho \geq 0.89$ [19], [24], [25], [26]. Hence, there is room for



(a) Attribution map: sensitivity analysis.



(b) Attribution map: Integrated Gradients, zero matrix baseline.

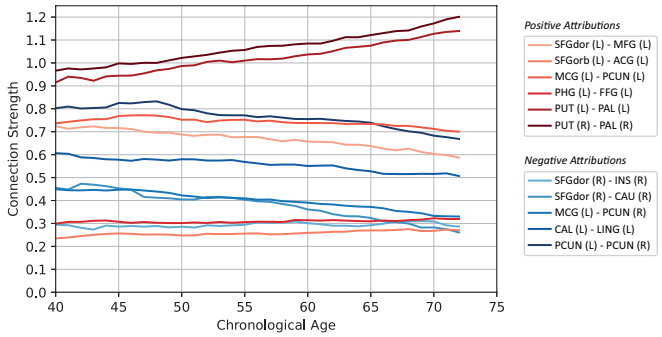


(c) Attribution map: Integrated Gradients, low prediction baseline.

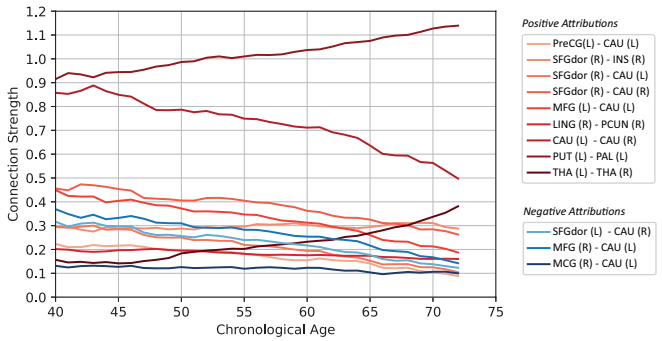
Fig. 7: Three different attribution maps are computed using sensitivity analysis (7a), Integrated Gradients with a zero matrix baseline (7b) and Integrated Gradients applied to the elderly, with the lowest predicted subject as baseline (7c). All three attribution maps are shown from an axial (left) and coronal (right) point of view. The blue edges represent connections that have a negative influence on the prediction score. Conversely, the red edges represent connections that positively influence the prediction score. The thickness of the edges corresponds to the absolute value of the gradient. The node size corresponds to the node strength which is calculated by taking the sum over the gradients per node. The attribution maps in 7a and 7b greatly overlap, whereas 7c shows more red and more interhemispheric edges.

TABLE II: An overview of the inter-subject and inter-model variation for all methods. The variation is expressed as the median coefficient of variation (CV) over all edges. Additionally, the median CV over the attributions of interest is shown.

		Sensitivity Analysis	Integrated Gradients *zero baseline*	Integrated Gradients *low prediction baseline*
All attributions	Inter-subject variation (CV)	2.15	1.69	1.90
	Inter-model variation (CV)	1.54	1.54	1.98
Attributions of interest	Inter-subject variation (CV)	0.622	0.517	1.16
	Inter-model variation (CV)	0.538	0.496	1.17



(a) Average connection strength for the connections found by the sensitivity analysis and Integrated Gradients with a zero matrix baseline.



(b) Average connection strength for the connections found by the Integrated Gradients with the lowest prediction subject as baseline.

Fig. 8: Average connection strength per age group for the connections corresponding to the attributions of interest. An extra threshold of 0.7 is applied on the absolute values of the attributions.

improvement in the performance of the proposed network and training procedure.

Firstly, a possibility to improve the performance of the network is to optimize the pre-processing of the data. The performance of the model is dependent on the quality of the input data. The AAL atlas used to parcellate the cerebrum is based on probablistic locations of the gyri and sulci [46] and linear co-registration is used to deform the dMRI data onto the atlas. However, more recent developments have shown inaccuracies in this method and improved the automated cortical parcellation [62].

Secondly, a major point of discussion is the ground truth of brain age. The training set is labeled using chronological age. Although the subjects are assumed to be healthy, it is not guaranteed that their chronological age matches their brain age. It is known that the used dataset is oversampled on type 2 diabetes and white matter structure has shown to be altered in type 2 diabetes patients compared to healthy controls [63], which might effect their brain age. Therefore, this could induce confusion in the training process of the model and/or lead to an overestimation of the mean absolute error on the test set. Future research could look into the improvement of the labels by strict selection on diabetes, CSVD (e.g. exclude all patients with a score of 1 or higher) and inclusion of e.g. cognitive performance.

Lastly, the scatter- and boxplot in Fig. 5 and 6 show a bias in the predictions on the test set towards the mean age. The explanation is two-fold: the mean squared error loss function forces the network to predict closer to the mean and this effect is enhanced by the non-uniform distribution of age in both the training and test cohort. Fig. 6 shows in orange numbers that the youngest and oldest age group are smaller in size and these groups have on average a larger mean absolute error. This effect is also observed by Han Peng et al. [26] who applied a bias correction as proposed by Smith et al. [64] to improve the performance of their network.

Due to the bias in the predictions, the training process needs improvement before brain age can be investigated. One of the main goals of brain age research is to identify factors that contribute to accelerated or delayed ageing. With the current study, the feasibility of brain age prediction on dMRI derived structural connectivity data has been shown. However, follow up research is necessary to correct for bias, improve the labelling of the training set and subsequently, to examine the subjects that have a large deviation in their chronological and brain age.

4.2 Explainable Deep Learning

As was discussed in the introduction, early research has shown that brain structures that myelinate last are more vulnerable to the effects of ageing. These structures include the association cortices and neostriatum, from which the latter consists of the caudate and putamen (Appendix A.3 further elaborates on the anatomy of the brain).

The edge between the putamen (PUT) and pallidum (PAL) is present in all three attribution maps, in both hemispheres (with the exception in Fig. 7c in the right hemisphere). These regions were also found by Bellantuono et al. as being predictive of age [19]. The nodes of the dorsolateral superior frontal gyrus (SFGdor) and medial frontal gyrus (MFG) are also present in all three attribution maps, in both hemispheres. The superior and medial frontal lobe appear to have a large contribution to the prediction score, which is in line with the statements of Raz who reports significant age-induced changes in the prefrontal cortex [8]. The MFG was also found to contribute to the prediction of age in previous research [19], [25], [65].

The edges between the SFGdor and the insula (INS) and the SFGdor and the caudate (CAU) in the right hemisphere

are present in all three attribution maps. However, with the sensitivity analysis and Integrated Gradients with a zero baseline, a negative contribution to the output was found, whereas in Fig. 7c the attributions are positive. This can be explained by equation 5, which shows that the integral is multiplied with the difference between input and baseline. Assuming that negative contributions represent connections that decrease in strength over age, the difference $(x_i - \bar{x}_i)$ for the young baseline and elderly subjects becomes negative and therefore, the final attribution becomes positive. Hence, the larger proportion of positive attributions in Fig. 7c in comparison with the attribution maps in Fig. 7a and 7b can be explained.

The sensitivity analysis and Integrated Gradients with a zero baseline resulted in very similar attribution maps. The similarity in the methods is the multiplication with the input itself (see equation 4 and 5). Because of the zero baseline, the difference between x_i and $\bar{x}_i$ equals x_i. Due to the large proportion of zeros in the input data (62%) many attributions are reduced to zero. This explains why the noise is reduced in the sensitivity analysis, but also raises the question if these gradients are indeed not important. The absence of a connection can be as valuable as its presence. Also, connections that are in general stronger than others have a higher attribution due to this multiplication, which might explain the strong presence of the PAL and PUT in Fig. 7a and 7b. Future research should investigate if the multiplication of the input itself does not influence the attributions too much, which could explain the similarities.

The Integrated Gradients method with a young subject as baseline on the other hand resulted in a very different attribution map. From a theoretical perspective, this method mimics the increase of the prediction score from 0 to 1 which corresponds with the original design of the Integrated Gradients method. The limitations are, however, that there is only one subject used as a baseline, which makes the method sensitive for noise and outliers. Also, fewer subjects were used to compute the attribution maps increasing the possible effect of noise and outliers. Table II shows that the variation in the attributions found by this last method is indeed larger compared to the other two methods. As stated previously, association tracts detoriate more than commissural tracts, which is in line with last-in-first-out hypothesis. Especially Fig. 7c shows strong edges within the interhemispheric (commissural) tracts, which contradicts this hypothesis.

The input features that are found to be most predictive of age, are evaluated via comparison with the average connection strength of the corresponding connections per age, which is plotted in Fig. 8. The goal of these figures is to verify if there is a change in connectivity strength throughout age. For most negative attributions in Fig. 8a, it can be observed that the corresponding connections indeed decrease in connection strength with increasing age. A decrease in connectivity strength can be explained, since white matter connections generally decrease in density and weight due to ageing [66]. The connections between the PUT and PAL in both hemispheres are increasing with increasing age, which is less intuitive. Furthermore, Fig. 8 confirms that these connections are relatively strong.

There is not a clear increase or decrease observed for all connections. For example, the connection between the SFGdor and the INS in the right hemisphere in Fig. 8a remains fairly constant. It is likely that the relation between the partial derivative and input feature is more complex, than assumed in this analysis. Furthermore, this analysis evaluates each input feature independently. However, the partial derivative does not only vary with x_i, but also with the values of other input features [27]. This interaction effect is not taken into account in the evaluation of the results.

The CV as reported in Table II shows large variability in each method. The attributions of interest have a standard deviation that is half the mean value. It should be taken into account that dMRI also has its limitations. For example, it is more difficult to track long-range connections, which can lead to an increase in variation in these connections. Furthermore, it is known that false positive connections can be detected using DTI, which could cause a high variation as well. Research has been done on the intra-subject variation of different pre-processing strategies for dMRI data. A CV of 0.45 was found for the strongest connections (top 20%) [67]. Accordingly, it can be argued that there is a level of uncertainty in each connectivity matrix, and it is acceptable that this level of uncertainty reappears in the attributions.

4.3 Future Research: The Reserve Hypothesis

An alternative explanation for the differences in the development of age-related effects is the reserve hypothesis. Reserve can be described as a buffer capacity for brain damage and dementia in later life [68]. The hypothesis has been coined to explain the heterogeneity in the association between brain damage and cognitive decline. Current evidence on the reserve hypothesis is mainly based on studies using confounded proxies, such as education and occupational complexity [69]. However, it is still unknown what the underlying mechanism of reserve is. What structural and functional changes occur in someone building up a reserve? Some studies investigated reserve using neural tissue markers based on functional connectivity of the brain [70], [71]. Although the sample size of these studies was small, the results provide an interesting basis for further research.

Within the current study the feasibility of training on structural connectivity data has been shown, including a method to explain the input features that are contributing the most to the output score. The promising results rise curiosity for other prediction tasks. Therefore, as an extension to this research, the prediction of education level using this set up was tested to investigate the underlying mechanism of the reserve hypothesis. An overview of the experiments and prediction results can be found in Appendix E. Unfortunately the network is not able to learn features that can predict education level in the training, validation and test set. A possible explanation is that the relation between education and structural connectivity is more complex in comparison with age. It can be doubted if education level is indeed a well defined proxy for reserve. Suggested solutions are to increase the complexity of the model, increase the size of the input data using multiple imaging modalities or to add information to the label such as occupation and age.

5 Conclusion

Within this study, brain connectivity in ageing is explored using XAI. It has been shown that the CNN is able to predict age based on structural connectivity data. The analysis of the input data indicates that multiple brain regions have a large contribution to the prediction of age. These regions include the superior and medial frontal lobe, which is in line with previous findings on significant age-induced changes in the prefrontal cortex. Furthermore, the putamen and pallidum, both structures of the basal ganglia, appear to have a large contribution to the prediction score as well as the precuneus. Further research is necessary to investigate the relation between the connections that are predictive of age and the ageing process. Additionally, further research should investigate the differences in the found attributions between the different XAI methods and the variation of the attributions within the methods.

References

[1] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013.

[2] Martin Prince, Renata Bryce, Emiliano Albanese, Anders Wimo, Wagner Ribeiro, and Cleusa P Ferri. The global prevalence of dementia: a systematic review and metaanalysis. *Alzheimer's & dementia*, 9(1):63–75, 2013.

[3] Theo Vos, Abraham D Flaxman, Mohsen Naghavi, Rafael Lozano, Catherine Michaud, Majid Ezzati, Kenji Shibuya, Joshua A Salomon, Safa Abdalla, Victor Aboyans, et al. Years lived with disability (ylds) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *The lancet*, 380(9859):2163–2196, 2012.

[4] Elaine N. Marieb and Katja N. Hoehn. *Human anatomy & physiology*. Pearson, ninth edition, 2014.

[5] Werner Kahle. *Atlas van de anatomie: Zenuwstelsel en zintuigen*. 2007.

[6] Junling Gao, Raymond Tak-Fai Cheung, Tatia Lee, Leung-Wing Chu, Ying-Shing Chan, Henry Ka-Fung Mak, John X Zhang, Deqiang Qiu, Germaine Fung, and Charlton Cheung. Possible retrogenesis observed with fiber tracking: an anteroposterior pattern of white matter disintegrity in normal aging and alzheimer's disease. *Journal of Alzheimer's Disease*, 26(1):47–58, 2011.

[7] Naftali Raz. Aging of the brain and its impact on cognitive performance: Integration of structural and functional findings. 2000.

[8] Naftali Raz. Ageing and the brain. *e LS*, 2001.

[9] Catherine Lebel and Sean Deoni. The development of brain white matter microstructure. *Neuroimage*, 182:207–218, 2018.

[10] Laurent Hermoye, Christine Saint-Martin, Guy Cosnard, Seung-Koo Lee, Jinna Kim, Marie-Cecile Nassogne, Renaud Menten, Philippe Clapuyt, Pamela K Donohue, Kegang Hua, et al. Pediatric diffusion tensor imaging: normal database and observation of the white matter maturation in early childhood. *Neuroimage*, 29(2):493–504, 2006.

[11] Po H Lu, Grace J Lee, Erika P Raven, Kathleen Tingus, Theresa Khoo, Paul M Thompson, and George Bartzokis. Age-related slowing in cognitive processing speed is associated with myelin integrity in a very healthy elderly sample. *Journal of clinical and experimental neuropsychology*, 33(10):1059–1068, 2011.

[12] Andrew R Bender, Manuel C Völkle, and Naftali Raz. Differential aging of cerebral white matter in middle-aged and older adults: a seven-year follow-up. *Neuroimage*, 125:74–83, 2016.

[13] John H. Martin. *Neuroanatomy*. 1989.

[14] James H Cole and Katja Franke. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*, 40(12):681–690, 2017.

[15] Katja Franke, Gabriel Ziegler, Stefan Klöppel, and Christian Gaser. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892, 2010.

[16] James H Cole, Stuart J Ritchie, Mark E Bastin, MC Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385–1392, 2018.

[17] Jason Steffener, Christian Habeck, Deirdre O'Shea, Qolamreza Razlighi, Louis Bherer, and Yaakov Stern. Differences between chronological and brain age are related to education and self-reported physical activity. *Neurobiology of aging*, 40:138–144, 2016.

[18] Eileen Luders, Nicolas Cherbuin, and Christian Gaser. Estimating brain age using high-resolution pattern recognition: Younger brains in long-term meditation practitioners. *Neuroimage*, 134:508–513, 2016.

[19] Loredana Bellantuono, Luca Marzano, Marianna La Rocca, Dominique Duncan, Angela Lombardi, Tommaso Maggipinto, Alfonso Monaco, Sabina Tangaro, Nicola Amoroso, and Roberto Bellotti. Predicting brain age with complex networks: From adolescence to adulthood. *Neuroimage*, 225:117458, 2021.

[20] World Health Organization. *World report on ageing and health*. WHO, Luxembourg, 2015.

[21] Katja Franke and Christian Gaser. Ten years of brainage as a neuroimaging biomarker of brain aging: what insights have we gained? *Frontiers in neurology*, 10:789, 2019.

[22] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[23] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.

[24] Gidon Levakov, Gideon Rosenthal, Ilan Shelef, Tammy Riklin Raviv, and Galia Avidan. From a deep learning model back to the brain—identifying regional predictors and their relation to aging. *Human brain mapping*, 41(12):3235–3252, 2020.

[25] Nicola Amoroso, Marianna La Rocca, Loredana Bellantuono, Domenico Diacono, Annarita Fanizzi, Eufemia Lella, Angela Lombardi, Tommaso Maggipinto, Alfonso Monaco, Sabina Tangaro, et al. Deep learning and multiplex networks for accurate modeling of brain age. *Frontiers in aging neuroscience*, 11:115, 2019.

[26] Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871, 2021.

[27] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Gradient-based attribution methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 169–191. Springer, 2019.

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[30] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[31] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[32] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[33] Chang-Le Chen, Yung-Chin Hsu, Li-Ying Yang, Yu-Hung Tung, Wen-Bin Luo, Chih-Min Liu, Tzung-Jeng Hwang, Hai-Gwo Hwu, and Wen-Yih Isaac Tseng. Generalization of diffusion magnetic resonance imaging–based brain age prediction model through transfer learning. *NeuroImage*, 217:116831, 2020.

[34] Owen Carmichael and Samuel Lockhart. The role of diffusion tensor imaging in the study of cognitive aging. *Brain Imaging in Behavioral Neuroscience*, pages 289–320, 2011.

[35] David J Madden, Ilana J Bennett, Agnieszka Burzynska, Guy G Potter, Nan-kuei Chen, and Allen W Song. Diffusion tensor imaging of cerebral white matter integrity in cognitive aging. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1822(3):386–400, 2012.

[36] Sune Nørhøj Jespersen. White matter biomarkers from diffusion mri. *Journal of Magnetic Resonance*, 291:127–140, 2018.

[37] Derek K Jones, Andrew Simmons, Steve CR Williams, and Mark A Horsfield. Non-invasive assessment of axonal fiber connectivity in the human brain via diffusion tensor mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(1):37–41, 1999.

[38] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[39] Miranda T Schram, Simone JS Sep, Carla J van der Kallen, Pieter C Dagnelie, Annemarie Koster, Nicolaas Schaper, Ronald MA Henry, and Coen DA Stehouwer. The maastricht study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. *European journal of epidemiology*, 29(6):439–451, 2014.

[40] Joost JA de Jong, Jacobus FA Jansen, Laura WM Vergoossen, Miranda T Schram, Coen DA Stehouwer, Joachim E Wildberger, David EJ Linden, and Walter H Backes. The effect of mr image quality on structural and functional brain connectivity: The maastricht study. *bioRxiv*, page 806075, 2019.

[41] Pinar Yilmaz, M Kamran Ikram, Wiro J Niessen, M Arfan Ikram, and Meike W Vernooij. Practical small vessel disease score relates to stroke, dementia, and death: the rotterdam study. *Stroke*, 49(12):2857–2865, 2018.

[42] AJBSJJDK Leemans, B Jeurissen, J Sijbers, and DK Jones. Exploredti: a graphical toolbox for processing, analyzing, and visualizing diffusion mr data. In *Proc Intl Soc Mag Reson Med*, volume 17, 2009.

[43] Jonathan AD Farrell, Bennett A Landman, Craig K Jones, Seth A Smith, Jerry L Prince, Peter CM Van Zijl, and Susumu Mori. Effects of signal-to-noise ratio on the accuracy and reproducibility of diffusion tensor imaging–derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5 t. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 26(3):756–767, 2007.

[44] Alexander Leemans and Derek K Jones. The b-matrix must be rotated when correcting for subject motion in dti data. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 61(6):1336–1349, 2009.

[45] Chantal MW Tax, Ben Jeurissen, Sjoerd B Vos, Max A Viergever, and Alexander Leemans. Recursive calibration of the fiber response function for spherical deconvolution of diffusion mri data. *Neuroimage*, 86:67–80, 2014.

[46] Edmund T Rolls, Marc Joliot, and Nathalie Tzourio-Mazoyer. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *Neuroimage*, 122:1–5, 2015.

[47] Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[48] Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

[49] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

[50] Colin J Brown, Jeremy Kawahara, and Ghassan Hamarneh. Connectome priors in deep neural networks to predict autism. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 110–113. IEEE, 2018.

[51] Achraf Essemlali, Etienne St-Onge, Maxime Descoteaux, and Pierre-Marc Jodoin. Understanding alzheimer disease's structural connectivity through explainable ai. In *Medical Imaging with Deep Learning*, pages 217–229. PMLR, 2020.

[52] François Chollet and contributors. Keras, 2015.

[53] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

[54] Vikas Verma, Meng Qu, Kenji Kawaguchi, Alex Lamb, Yoshua Bengio, Juho Kannala, and Jian Tang. Graphmix: Improved training of gnns for semi-supervised learning. *Preprint arXiv:1909.11715*, 2019.

[55] Kezhi Kong, Guohao Li, Mucong Ding, Zuxuan Wu, Chen Zhu, Bernard Ghanem, Gavin Taylor, and Tom Goldstein. Flag: Adversarial data augmentation for graph neural networks. *Preprint arXiv:2010.09891*, 2020.

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

[57] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[58] Amitojdeep Singh, Sourya Sengupta, and Vasudevan Lakshminarayanan. Explainable deep learning models in medical image analysis. *Journal of Imaging*, 6(6):52, 2020.

[59] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 2017.

[60] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

[61] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7):e68910, 2013.

[62] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[63] Yael D Reijmer, Manon Brundel, Jeroen De Bresser, L Jaap Kappelle, Alexander Leemans, Geert Jan Biessels, Utrecht Vascular Cognitive Impairment Study Group, et al. Microstructural white matter abnormalities and cognitive functioning in type 2 diabetes: a diffusion tensor imaging study. *Diabetes care*, 36(1):137–144, 2013.

[64] Stephen M Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E Nichols, and Karla L Miller. Estimation of brain age delta from brain imaging. *Neuroimage*, 200:528–539, 2019.

[65] Benson Mwangi, Khader M Hasan, and Jair C Soares. Prediction of individual subject's age across the human lifespan using diffusion tensor imaging: a machine learning approach. *Neuroimage*, 75:58–67, 2013.

[66] Richard F Betzel, Lisa Byrge, Ye He, Joaquín Goñi, Xi-Nian Zuo, and Olaf Sporns. Changes in structural and functional connectivity among resting-state networks across the human lifespan. *Neuroimage*, 102:345–357, 2014.

[67] Shang-Yueh Tsai. Reproducibility of structural brain connectivity and network metrics using probabilistic diffusion tractography. *Scientific reports*, 8(1):1–12, 2018.

[68] Yaakov Stern. Cognitive reserve. *Neuropsychologia*, 47(10):2015–2028, 2009.

[69] Linda Clare, Yu-Tzu Wu, Julia C Teale, Catherine MacLeod, Fiona Matthews, Carol Brayne, Bob Woods, and CFAS-Wales Study Team. Potentially modifiable lifestyle factors, cognitive reserve, and cognitive function in later life: A cross-sectional study. *PLoS medicine*, 14(3):e1002259, 2017.

[70] Marco Bozzali, Claire Dowling, Laura Serra, Barbara Spano, Mario Torso, Camillo Marra, Diana Castelli, Nicholas G Dowell, Giacomo Koch, Carlo Caltagirone, et al. The impact of cognitive reserve on brain functional connectivity in alzheimer's disease. *Journal of Alzheimer's Disease*, 44(1):243–250, 2015.

[71] Nicolai Franzmeier, MÁ Araque Caballero, ANW Taylor, Lee Simon-Vermot, K Buerger, Birgit Ertl-Wagner, Claudia Mueller, Cihan Catak, Daniel Janowitz, Ebru Baykara, et al. Resting-state global functional connectivity as a biomarker of cognitive reserve in mild cognitive impairment. *Brain imaging and behavior*, 11(2):368–382, 2017.

Appendix A
Neuroanatomy and Terminology

A.1 Terminology

Functional Connectivity = The temporal dependency of neuronal activation patterns of anatomically separated regions [73].

Lacunar infarcts = Small non-cortical infarcts that result from occlusion of penetrating branches of the large cerebral arteries [74].

Microbleeds = A (cerebral) microbleed is a chronic accumulation of small blood products in the brain tissue [75].

Structural connectivity = The physical white matter tracts that connect anatomically separated regions [73].

White matter hyperintensities = Lesions mainly caused by demyelination and axonal loss as a consequence of chronic ischaemia. The prevalence and severity of these lesions increase with age [76].

A.2 Deep Learning Terminology

This section contains the explanation of terms that are used to describe the architecture and training process of the deep learning network. The definitions are based on Goodfellow et al. [22].

Activation function = A nonlinear function applied to each neuron in a deep learning network. This function determines the output of that neuron for a given input.

Batch size = The number of training samples in a single iteration.

Dropout = A regularization technique that randomly excludes a defined percentage of the neurons from the training process.

Generalization error = The expected value of the error on an unseen dataset, also called the test error.

Hyperparameters = Settings within the deep learning network that are used to control the behaviour of the learning algorithm.

Learning rate = A positive scalar that determines the step size of the optimization algorithm.

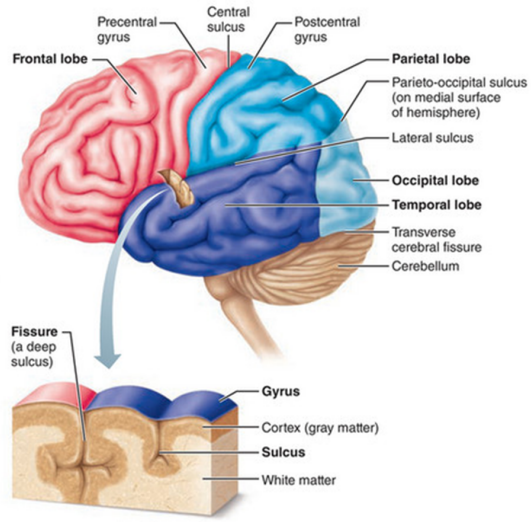
Loss function = The function that the learning algorithm wants to minimize.

Optimizer = An algorithm that is designed to find the parameters of a neural network that significantly reduce the loss function.

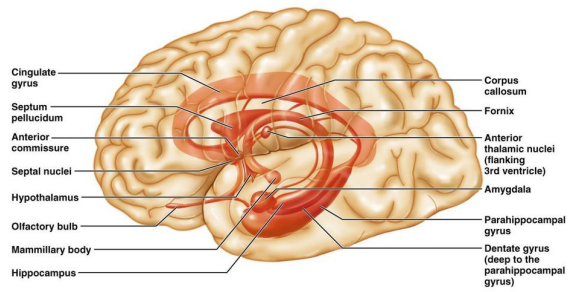
Regularization = Any modification to the learning algorithm to reduce the generalization error, but not its training error.

A.3 Neuroanatomy

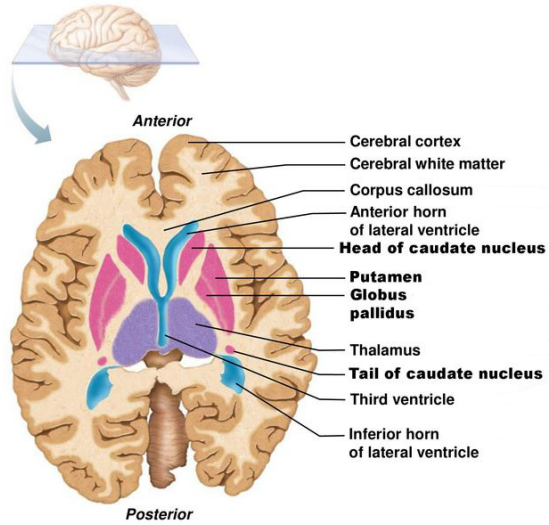
This section contains three images to illustrate different parts of the brain, see Fig. A.1.



(a) Schematic view of the brain showing the four main lobes of the cerebral hemispheres. The frontal lobe is important for cognitive behaviour and the control of voluntary movement. The parietal lobe contains the primary and association sensory areas. The occipital lobe processes visual information and the temporal lobe processes memories and auditory information.



(b) The limbic system is the emotional brain. It mediates the emotional response and is involved in memory processing. Lower and higher brain regions are extensively connected to the limbic system, which allows the system to interact with a variety of stimuli.



(c) The basal ganglia are a group of subcortical nuclei, which includes the caudate, the putamen, and the globus pallidus or pallidum. These nuclei receive input from all four lobes and help control skeletal muscle movements. In addition, they play a role in emotion and cognition.

Fig. A.1: Anatomy of the brain: the four lobes (a), the limbic system (b) and the basal ganglia (c). Images and information are from [4].

Appendix B
Experiments on hyperparameters

This section contains a brief overview of the experiments that are conducted to achieve an optimal training procedure. Inspired by the publication on No new-net by Isensee et al. the focus within these experiments is on the training procedure rather than the network architecture [77]. The training cohort is split into a train (60%), validation (20%) and test (20%) set to conduct the experiments. The experiments are evaluated based on the resulting loss curves and predictions on the test set.

The first experiments are set up as a binary classification problem. The predictions on the test set can be true positive (TP), false postive (FP), true negative (TN) or false negative (FN). The experiments are evaluated using the accuracy of the predictions on the test set which is calculated according to equation B.1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{B.1}$$

Later experiments address brain age prediction as a regression problem and are evaluated using the mean absolute error as previously described in section 2.4.3.

B.1 Classification Versus Regression

The experiments started with BrainNetCNN as a classification model. The subjects were divided over a young and old class based on the median of the age distribution to generate two equally sized classes. The loss function used is binary cross entropy. Fig. B.1a shows the loss curve. An accuracy of 75.6% was achieved on the test set.Variation in the architecture gave similar results with an maximum accuracy of 78%.

It was hypothesized that the performance could not further increase because subjects aged close to the median are difficult to classify. Therefore an additional experiment was conducted on the extremes only. Two classes were formed: $40 - 50$ years and $68 - 79$. The age limits were chosen such that the classes were equal in size. The loss curve is shown in Fig. B.1b and the accuracy on the test set was 98.6%.

The transfer to a regression problem with the mean squared error as a loss function resulted in more smoothly declining loss curves. The idea of training on extremes was used to experiment with transfer learning. The model was first trained on the extreme ages first as defined previously, which resulted in a mean absolute error (MAE) of 3.50 years and the model was trained a second time on all subjects. The resulting MAE was 5.57 years and the loss curve is in Fig. B.1c.

Within the training procedure there - Mean squared error penalizes large errors more than the mean absolute error function would. A experiment was conducted in which the network was trained using the mean absolute error loss function to confirm this hypothesis. Alternatively, the loss can be improved by optimization of the loss function as suggested by Brown et al. [50].

B.2 Regularization

As proposed by Kawahara et al. and Essemlali et al., the network is trained using L2 regularization [49], [51]. This is used to reduce the size of the weights and to prevent overfitting. Fig. B.2a shows the weights of the network while training as proposed in section 2.4. Fig. B.2b on the other hand, shows the weights of the same network, however, it is now trained without L2 regularization. The effect of L2 regularization on the size of the weights can be clearly observed from the width of the histograms in these figures. Additionally, L2 is added to the training procedure with a weight decay of 0.0001, which is a factor 5 smaller in comparison with the proposed training procedure. The resulting weights are plotted in Fig. B.2c. The dropout rate is increased to 0.5 to prevent overfitting in the training procedures without L2 regularization and with reduced L2 regularization. The performance of the training without L2 regularization improved, in comparison with the proposed procedure. The mean absolute error on the test set reduced with 0.3 years. However, the network started to overfit after 30 epochs. Therefore, it is decided to train with L2 regularization.

B.3 Batch Size

BrainNetCNN was first developed and trained with a batch size of 14 [49]. Follow up research on the BrainNetCNN fine-tuned the architecture which resulted in a bacth size of 25. With this as reference, the network was trained using a batch size of 14, 25, 45, 65 and 85.The results are given in Table B.I.

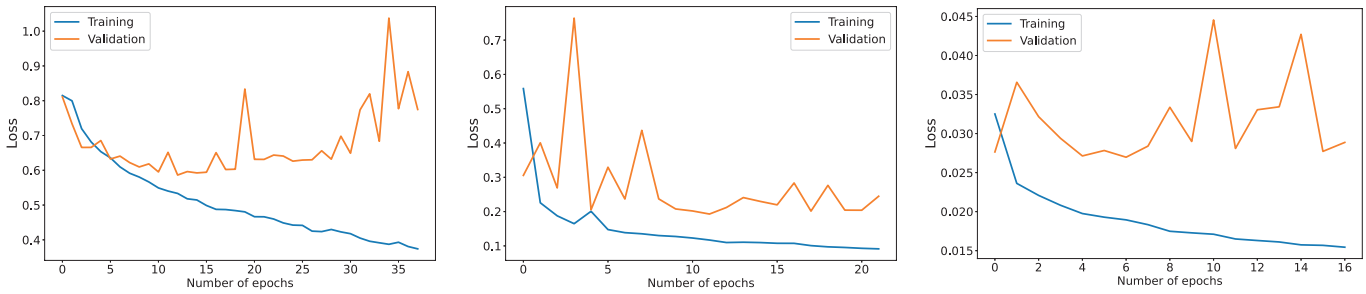
TABLE B.I: The results on a single training with varying batch size. The test set is a subset of the training cohort that was not used in the training itself.

Batch Size	Test MAE	Test ρ
14	5.42	0.638
25	5.27	0.666
45	5.21	0.653
65	5.55	0.623
85	5.60	0.628

Appendix C
Experiments on Data Augmentation

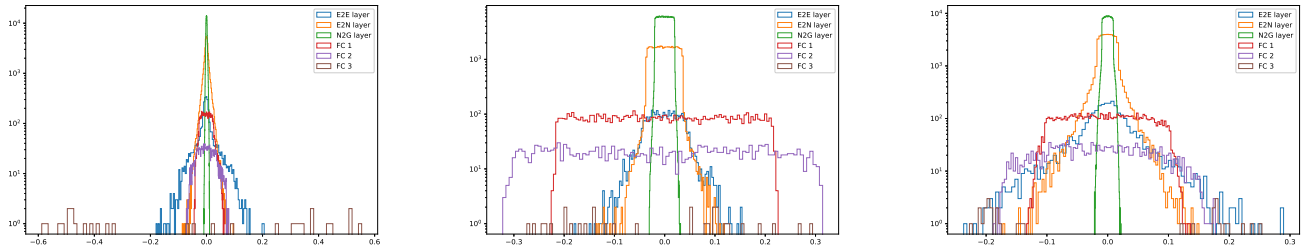
As described previously, data augmentation is a well-known regularization technique. However, label-preserving transformations are not applicable to graph networks. Therefore, experiments have been conducted to investigate different data augmentation techniques for structural connectivity data. The experiments and corresponding results to test the effect of the augmentation methods are reported in this section.

Four different augmentation methods have been tested, including the Keras Gaussian Noise layer. Their contribution to the performance of the network is evaluated using 5-fold cross validation. Three out of four methods are based on the addition of noise to the original data. The scale of noise is determined using the data distribution of all connectivity matrices, which is shown in the histogram in Fig. C.1. From this figure, it can be observed that the majority of the non-zero data points are smaller than 0.01.



(a) (b) (c)

Fig. B.1: The training and validation loss curves for three different training procedures. From left to right: a classification network trained to predict low or high age (a), a classification network trained on the youngest (40 − 50 years) and oldest subjects (68 − 79) only (b), a regression network initialized with pre-trained weights (c).



(a) (b) (c)

Fig. B.2: The histograms of the weights of each layer for three different training procedures. From left to right: the weights the network using the proposed architecture (a), the weights of the network, while training without L2 regularization, and with a dropout of 0.5 (b), the weights of the network, while training with a L2 regularization weight decay of 0.0001, and with a dropout of 0.5. The effect of L2 regularization can be observed from the figures.

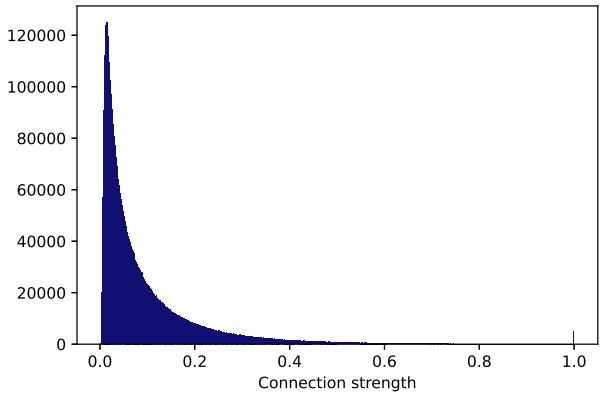


Fig. C.1: The distribution of the connectivity values in the training set. This histogram does not contain the zero values which make up 62% of all the data points.

C.1 Addition of noise through random re-scaling

Comparable to the Keras Gaussian Noise layer, this method adds random noise to each edge to generate more input data. Here, random relative noise is added to each edge before the training process starts. The main advantage of relative noise is that zero elements remain zero, and only positive values are generated. To add circa 10% of noise, a random noise matrix $\mathbf{N}$ is computed with elements drawn from a normal distribution with mean $\mu = 1$ and standard deviation $\sigma = 0.1$. This matrix is multiplied elementwise (the Hadamard product) with the connectivity matrix $\mathbf{A}$. To retain a symmetric matrix the lower half of the matrix is mirrored to the upper half.

C.2 Addition of noise based on edge variation

Alternatively, the amount of noise can be based on the standard deviation σ_{ij} of edge ij. The idea is to add noise to each edge based on the average variation of that edge. Additionally, the standard deviation matrix $\mathbf{S}$ is weighted by a random scalar $\lambda \in [0, 2]$ to add noise within two standard deviations, as shown by Equation C.1.

$$\mathbf{A} = \mathbf{A} + \lambda \mathbf{S}_a \tag{C.1}$$

With $\mathbf{S}_a$ the element-wise standard deviation matrix for age a. All adjacency matrices are symmetric, therefore $\mathbf{S}_a$ is also symmetric.

C.3 Simulation of synthetic connectivity matrices

One of the challenges in data augmentation is to apply low-cost transformations that preserve the semantic meaning of the graphs. In this section a data augmentation method is proposed that generates synthetic samples to enlarge the training set and to create equally sized classes. Generation of synthetic samples has shown its benefits in previous research [49],[78].

Based on these references it can be assumed that structural connectivity is similar among individuals with the

same chronological age (or education class). Furthermore, by swapping regions it is assumed that the model is forced to ignore confounders such as gender and diabetes mellitus.

Within a class, two random matrices $\mathbf{A}_m$ and $\mathbf{A}_n$ are selected and a number of regions from matrix $\mathbf{A}_m$ are replaced with the same regions from $\mathbf{A}_n$. The regions are randomly chosen and the number of regions replaced is 10, 20, 30 or 40. As described by Chawla et al., it is expected that a larger and less specific decision space is created, which results in a network that generalizes better [78]. To generate even more diversity within the synthetic samples, this augmentation method is applied in combination with the Keras Gauss noise layer.

C.4 Experiments

In this research four different augmentation methods are tested and their contribution to the performance of the network is evaluated using 5-fold cross validation. To evaluate the augmentation methods the evaluation metrics are calculated for each trained model, while predicting age for an independent test set on which no augmentation is applied. The metrics are averaged over 5 bootstrapped test samples of size 300 (with replacement). In addition, the ensemble performance of each method is computed. The models that are evaluated are as follows:
1. Baseline model
2. Baseline model + Gauss
3. Baseline model + random re-scaling
4. Baseline model + std based noise
5. Baseline model + synthetics + Gauss

The training set contains almost 2000 participants, which is already a relative large set compared to other medical image datasets. To evaluate the impact of data augmentation, a augmentation factor of 6 is used. It is assumed to be enough to measure a difference, without leading to overfitting as the data becomes too similar. However, it is still arbitrarily chosen and therefore, it is suggested that this factor is tested in future research. This means that for every matrix in the training and validation set, 6 new matrices are generated. For model 5, synthetic matrices are created that scales up the dataset to a factor 6, and in addition, obtains equal classes.

C.5 Results

The results of the data augmentation experiments are summarized in Table C.I. According to these results, the best performing model is the base model in combination with the Keras Gaussian Noise layer.

TABLE C.I: The results on five bootstrapped test sets for four different data augmentation methods. The mean absolute eror (MAE) and Pearson's correlation coefficient (ρ) are given to evaluate the performance of each cross-validation.

	Test MAE (years)	Test ρ (-)
1. Base Model	5.55	0.611
2. Base Model + Gaussian Noise	5.40	0.636
3. Base Model + Random Rescaling	5.66	0.606
4. Base Model + STD based noise	5.56	0.621
5. Base Model + Synthetics + Gaussian Noise	5.44	0.628

Appendix D
Additional Results

D.1 Loss Graphs

The training and validation loss curve of the prediction of brain age in Fig. D.1 rapidly decreases within the first 100 epochs and then converge. The training curves approach a loss of 0.01 and the validation curves approach 0.04.

D.2 Coefficient of Variation per Connection

The average CV per connection for the attributions of interest is reported in Table D.I. Note that the attributions of interest do not overlap for all three methods. Therefore, some rows of the table are partly empty.

Appendix E
Training on Education Level

As an extension to this research, the prediction of education level using this set up was tested to investigate the underlying mechanism of the reserve hypothesis. An overview of the experiments and prediction results are reported in this section.

Experiments have been performed using the input data and model described in this report to predict the level of education of the subjects in the Maastricht study. The level of education is divided into three categories (approximately in tertiles): low (no education, (un)completed primary education, or lower vocational education), medium (intermediate vocational education or higher secondary education) and high (higher vocational education or university education). The medium category was omitted from the first experiments to simplify the prediction task.

E.1 Network architecture

The input data is of the same format as described in section 2.2. The network is adapted towards a classification network which is previously done by Essemlali et al. [51]. The architecture and hyperparameter settings are adapted to their method. The N2G convolution layer is removed and the amount of filters are further reduced. The first E2E convolution layer has 23 filters, the E2N layer has 59 filters and both FC layers have 19 filters. After each layer a batch normalization layer is added for better convergence during training. Each layer uses the leaky rectified linear units (ReLU) as an activation function with a slope of 0.016 for all values < 0.

E.2 Training Procedure

The loss function is replaced by the binary cross entropy loss function and the softmax activation function is added to the output neuron. The network is asked to predict in range $[0, 1]$ in which 0 corresponds to the low education class and 1 corresponds to the high education class. A validation (20%) and test set (20%) are created within the training cohort. The test set is defined prior to the training process and remains the same for all experiments. An epoch is defined as an iteration over all batches and the network is trained for a maximum of 1000 epochs with a batch size of

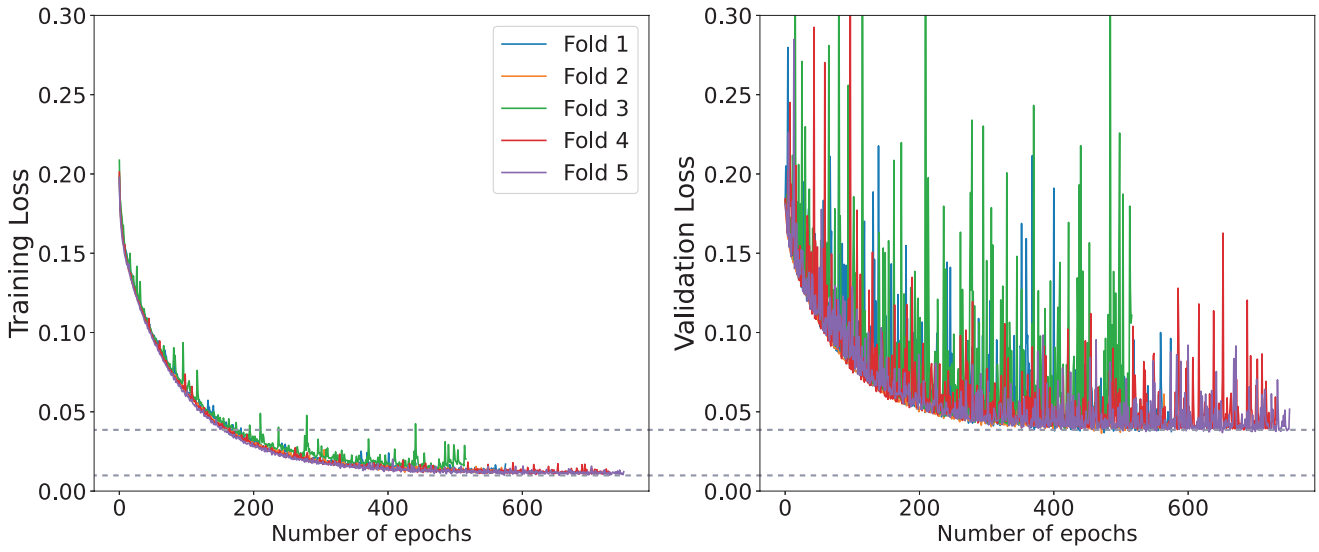


Fig. D.1: The loss graphs of the training process of each network. Left are the training loss curves shown and right the validation loss curves. The gray dotted lines visualize the difference between the final training and validation loss.

25. Training is terminated early if the validation loss has not improved within the last 30 epochs. Training is done using the Adam optimizer with an initial learning rate of 0.0078. An additional term in the loss function, the L2 regularization forces the trainable parameters to stay small. The weight decay used is 0.0005.

An alternative loss function for a binary classification problem is the hinge loss function. It was originally designed for support vector machines and is an widely used alternative to binary cross entropy for classification [79]. It requires a prediction score in range $[-1, 1]$ and a tahn activation function at the output neuron. A score of -1 represents the low education class and a score of 1 still represents the high education class.

Additionally, several hyperparameters are varied to investigate their influence on the training process. Each variation in the hyperparameter settings of the training procedure is tested in a single training. The training is evaluated based on the training and validation loss curve and the accuracy of the predicitions on the test set.

E.3 Results

The resulting loss curves of the training using binary cross entropy as a loss function as described in section E.2 can be found in Fig. E.1a. The validation loss starts to increase after 5 epochs which is an indication that the network is overfitting on the training data. The accuracy of the predictions on the test set is 56%. A second training is performed using the stochastic gradient descent optimizer instead of Adam which resulted in the loss curves in Fig. E.1c and an accuracy of 57% on the test set. The validation loss neither increases or decreases. This suggests that the validation set is not representative for the training set.

The training using the hinge loss function as described in section E.2 resulted in the training and loss curve as shown in Fig. E.1b and an accuracy of 51% on the test set. The gap between the training and validation loss again indicates overfitting on the training set. The training procedure using the hinge loss function is repeated, but terminated early if the validation loss has not improved within the last 100 instead of 30 epochs. The resulting loss curves can be found in Fig. E.1d. The training and validation loss decrease slightly within the first 100 epochs and then both reach a plateau. The accuracy on the test set improved to 55%.

More experiments were conducted varying the learning rate, batch size and dropout rate. However, no combination of hyperparameters was found that led to a significant improvement in the training process.

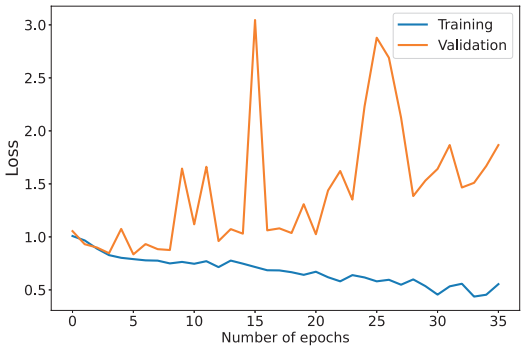
E.4 Discussion

The loss curves in Fig. E.1 show that the training does not converge. The training loss curve of the prediction of brain age in Fig. D.1 rapidly decreases in the beginning and then converges towards zero. Also the difference between the final training and validation loss is relatively small. This is not the case for the prediction of education level.

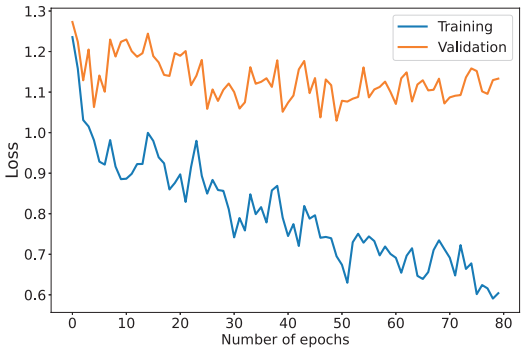
It appears that the model is not able to use the learned information from the training set to predict on the validation set. A possible explanation is that the relation between education and structural connectivity is more complex in comparison with brain age.A possible solution can be to increase the complexity of the network by adding an extra E2E layer, increasing the amount of filters or adding more fully connected layers. Alternatively, Graph Neural Networks (GNNs) could be investigated as they are especially designed for graph networks. Within the first experiments of the prediction of education level structural structural connectivity data was used, derived from dMRI. Previous studies investigated reserve based on functional connectivity of the brain [70], [71]. Within the Maastricht Study both structural and functional structural connectivity data is available. Functional connectivity is derived from resting-state functional MRI (rs-fMRI). In rs-fMRI the blood-oxygen-level-dependent (BOLD) signals are measured, thereupon

TABLE D.I: An overview of the inter-subject and inter-model variation per attribution of interest, for all methods. The variation is expressed as the average coefficient of variation (CV) over all subjects.

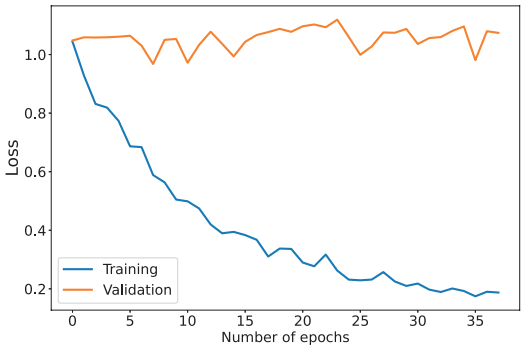
	Sensitivity Analysis		Integrated Gradients *zero baseline*		Integrated Gradients *low prediction baseline*	
Connection	Subject CV	Model CV	Subject CV	Model CV	Subject CV	Model CV
PreCG (L) - CAU (L)					0.87	1.81
PreCG (R) - PoCG (R)	-0.77	-0.55	-0.57	-0.53		
SFGdor (L) - MFG (L)	1.11	0.54	0.76	0.54		
SFGdor (L) - SFGmed (L)	-0.51	-0.87			-1.05	0.96
SFGdor (L) - CAU (R)					1.23	-1.54
SFGdor (R) - INS (R)	-0.88	-0.41	-0.73	-0.37	2.77	0.69
SFGdor (R) - CAU (L)					0.21	1.25
SFGdor (R) - CAU (R)	-3.55	-0.60	-1.46	-0.93	2.71	-3.51
MFG (L) - CAU (L)					0.82	0.35
MFG (R) - IFGtriang (R)	0.58	0.27	0.46	0.28		
MFG (R) - CAU (L)					-1.18	-6.84
ROL (L) - INS (L)			0.62	0.35		
ROL (R) - INS (R)	1.22	0.63	1.16	0.71		
SFGmed (R) - MCG (R)					-2.27	-0.83
SFGorb (L) - ACG (L)	0.56	0.32	0.46	0.42		
MCG (L) - PCUN (L)	0.78	0.30	0.47	0.25		
MCG (L) - PCUN (R)	-0.58	-0.44	-0.45	-0.37		
MCG (R) - CAU (L)					-1.50	-1.13
HIP (R) - PUT (R)					-12.48	2.07
PHG (L) - FFG (L)	0.58	0.22	0.46	0.17		
AMYG (L) - THA (R)					0.66	0.84
CAL (L) - LING (L)	-0.54	-0.29	-0.41	-0.20		
CUN (L) - SOG (L)	-0.47	-1.08	-0.21	-1.05		
LING (R) - PCUN (R)					9.60	1.11
SPG (L) - PCUN (L)	-0.38	-0.98				
PCUN (L) - PCUN (R)	-2.35	-0.74	-1.09	-0.65	5.89	2.23
CAU (R) - PAL (R)					1.14	0.39
CAU (R) - THA (R)	0.41	0.89				
PUT (L) - PAL (L)	2.04	0.54	0.68	0.54	0.95	0.83
PUT (R) - PAL (R)	0.58	0.49	0.37	0.43		
THA (L) - THA (R)					-0.38	2.66



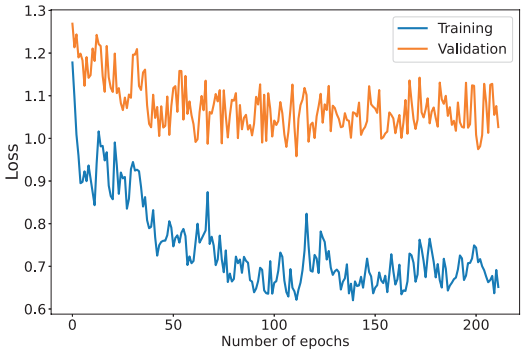
(a) The training and validation loss curve of training with the binary cross entropy loss function. An accuracy of 56% is achieved on the test set.



(b) The training and validation loss curve of training with the hinge loss function. An accuracy of 51% is achieved on the test set.



(c) The training and validation loss curve of training with the binary cross entropy loss function and stochastic gradient descent (SGD) as optimizer. An accuracy of 57% is achieved on the test set.



(d) The training and validation loss curve of training with the hinge loss function. The training is terminated if the validation loss does not change over the last 100 epochs. An accuracy of 55% is achieved on the test set.

Fig. E.1: The loss curves of training on dMRI data to predict education level

functional connectivity is determined by cross-correlating temporal changes in anatomically separated brain regions [38], [40],[80]. From this a connectivity matrix can be computed and used as input for the proposed deep learning network. A few experiments have been conducted using functional connectivity, however the resulting loss graphs were very similar to the graphs in Fig. E.1. The highest accuracy that is achieved using functional connectivity is 60%. This is achieved with training with a hinge loss function and Adam optimizer, a learning rate of 0.007 and dropout is omitted. The loss of the validation set did not decrease and its curve is similar to Fig. E.1d.

Lastly, it can be doubted if level of education is a well defined proxy for reserve. Especially subjects aged 40 - 77 years have had many years after their primary education that could have influenced their structural connectivity. Factors such as occupational complexity are expected to have an impact as well. A suggested solution is therefore to redefine the label of the data by including e.g. occupation and age.

Appendix F
Additional References

[73] Olaf Sporns. Structure and function of complex brain networks. *Dialogues in clinical neuroscience*, 15(3):247, 2013.

[74] C Miller Fisher. Lacunar strokes and infarcts: a review. *Neurology*, 32(8):871–871, 1982.

[75] G Roob, R Schmidt, P Kapeller, A Lechner, H-P Hartung, and F Fazekas. Mri evidence of past cerebral microbleeds in a healthy elderly population. *Neurology*, 52(5):991–991, 1999.

[76] Niels D Prins and Philip Scheltens. White matter hyperintensities, cognitive impairment and dementia: an update. *Nature Reviews Neurology*, 11(3):157–165, 2015.

[77] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. No newnet. *International MICCAI Brainlesion Workshop*, pages 234–244. Springer, 2018.

[78] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[79] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. *In Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[80] KJ Friston, CD Frith, PF Liddle, and RSJ Frackowiak. Functional connectivity: the principal-component analysis of large (pet) data sets. *Journal of Cerebral Blood Flow & Metabolism*, 13(1):5–14, 1993.