MA679 Final Project

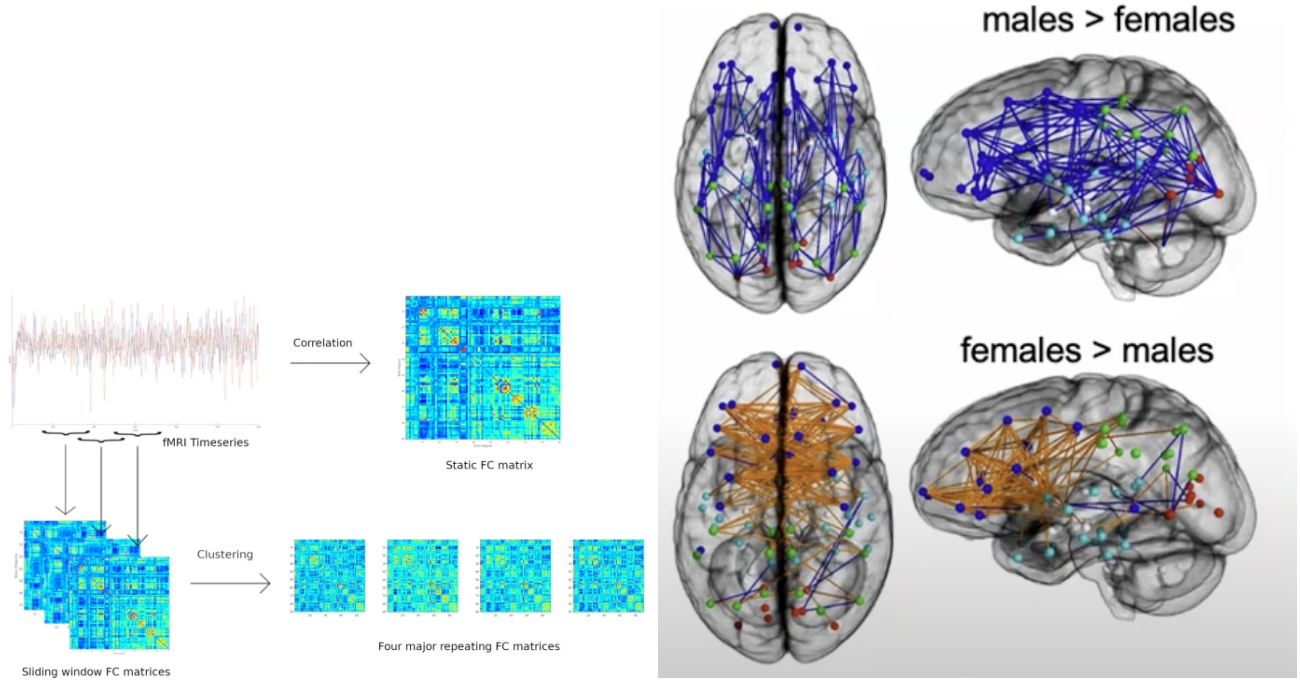
Suheng Yao, Ruijian Lin, Truc Minh Nguyen

May 5, 2025

1 Abstract

The motivation of this project is to design a model that predicts age using resting state data obtained from Functional Magnetic Resonance Imaging (FMRI). Through the utilization of appropriate machine learning techniques, we aim to identify key predictors of age-related anatomical changes to improve model accuracy, allowing for a meaningful comparison of male and female brain development. The relevant predictors will drive further biological research into the specific brain regions that capture the most variability in the response variable. A Linear Regression was included as a baseline reference to compare to more complex models such as Lasso, Ridge, Elastic Net, XGBoost, and Random Forest to analyze the bias-variance tradeoff. The interpretation of the model results serves as insight for clinicians to prepare proactive care for patients who are at risk of developing mental health issues.

2 Introduction



(a) Functional Connectome Matrix from fMRI (b) Functional Connectivity Comparisons

Figure 1: How to Get Connectome Matrix and Comparison between Male and Female

The fMRI data, obtained from the Healthy Brain Network (HBN), divides the brain into regions of interest, and extracts the average activity of each region over time. These regional time series are used to create 2 dimensional functional connectome matrices, where each entry represents the correlation between any two brain regions' time series. Larger positive values indicate stronger functional connectivity and higher similar time series between specified regions. Figure 1(a) above displays the process of obtaining the functional connectome matrix from fMRI time series data (Menon and Krishnamurthy 2019). The HBN dataset includes roughly 1,578 observations (aged 5-21), with an estimated sex proportion of 63% male and 37% female. Each individual has a functional connectivity network matrix (200×200), and additional information about the individual such as sex, ethnicity, race, etc.

The important biological concepts to present are the sex-specific differences in

healthy and disordered brains. On average, females have larger frontal gray matter and males have larger temporal gray matter. Females have stronger white matter functional connectivity between hemispheres (shown by the orange lines in Figure 1(b) above (WiDS Worldwide 2025)) and males have stronger connections within hemispheres (blue lines). Functional connectivity is the coactivation pattern of brain activity over time, the more similar the regions, the stronger the bond. Therefore, functional connectivity matrices are utilized to classify the sex of the brain.

3 Data Preprocessing and EDA

3.1 Data Cleaning

The training dataset contains data for 1104 participants, and the testing dataset contains data for 474 participants. Only the training dataset will be focused on during model fitting to avoid data leakage, and the testing dataset is only used for evaluation. For the functional connectivity network matrix, there were no missing values for any observations, therefore, the raw data contained all complete cases. The dataset was pivot wide, where each row is one observation containing the participant's ID and all the corresponding predictor variable values. There were missing values for the demographic variables, and the methods of imputation varied slightly depending on the model selections. The imputation method that provided the best performance was selected for each model. It is important to note that imputation is an estimation, and there is a margin of error that is associated with the results, which must be taken into context during interpretation. Principal Components Analysis (PCA) was also performed to reduce the dimension of the data while preserving content. The number of principal components selected for each model varied, depending on the PCA methods that gave the selected model the best performance.

3.2 Exploratory Data Analysis

All the Exploratory Data Analysis (EDA) is performed on the training dataset. The data contains roughly twice as many male samples as females; however, the age distribution of the training dataset appears to be mostly balanced between sexes, with one outlier in the female group (see Figure 2 below). The median age seems to be around 10 - 11 years, and the majority of the interquartile ranges appear to be around 8 - 14 years. Overall, the age distribution seem to be a little bit right skewed (see Figure 3 below), indicating there might need some log or power transformation of the response variable. In both datasets, the race sample count from highest to lowest is Whites, Others, Blacks, and Asians. However, the age distribution across different races is generally balanced (see Figure 4 below).

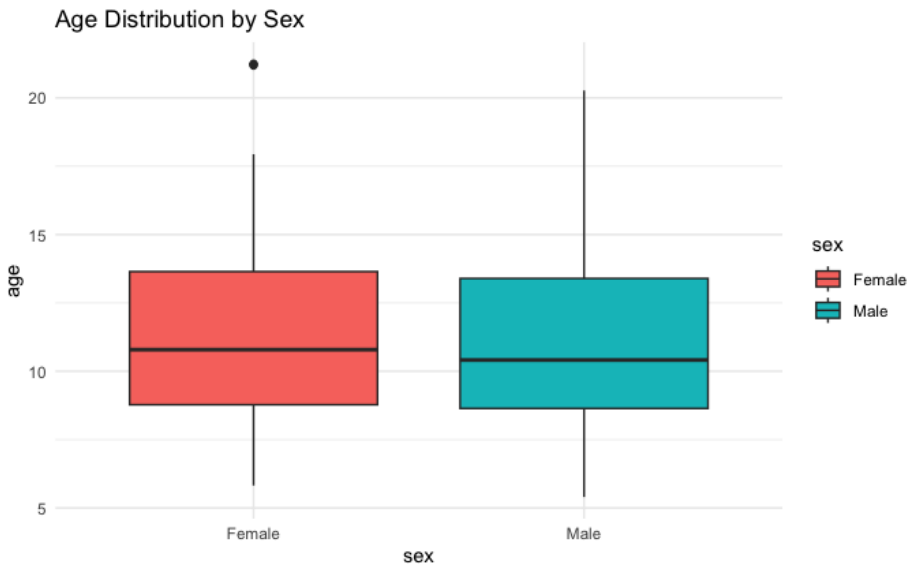


Figure 2: Age Distribution by Sex

Pearson Correlation calculations were performed to determine the top 50 functional connections correlated with age, which have the potential to be relevant covariates in the models to be explored. Functional connectome heatmaps were created to visualize the density of the correlations. It appears that the average functional connectome heatmap for females has higher warm-tones density (stronger correlation) than males, suggesting that, on average, females may exhibit stronger inter-regional synchronization during resting state fMRI (shown in

Figure 3: Overall Age Distribution

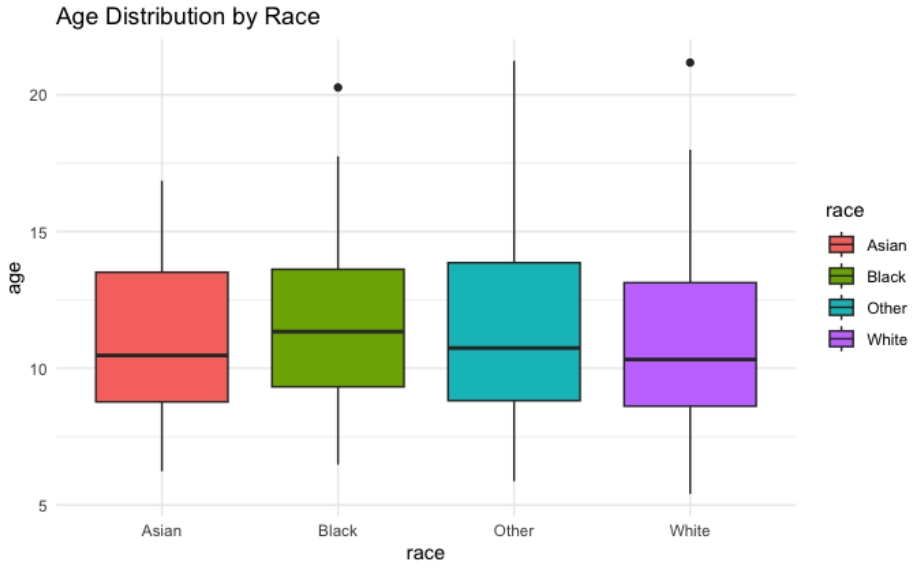


Figure 4: Age Distribution by Race

Figure 5 below).

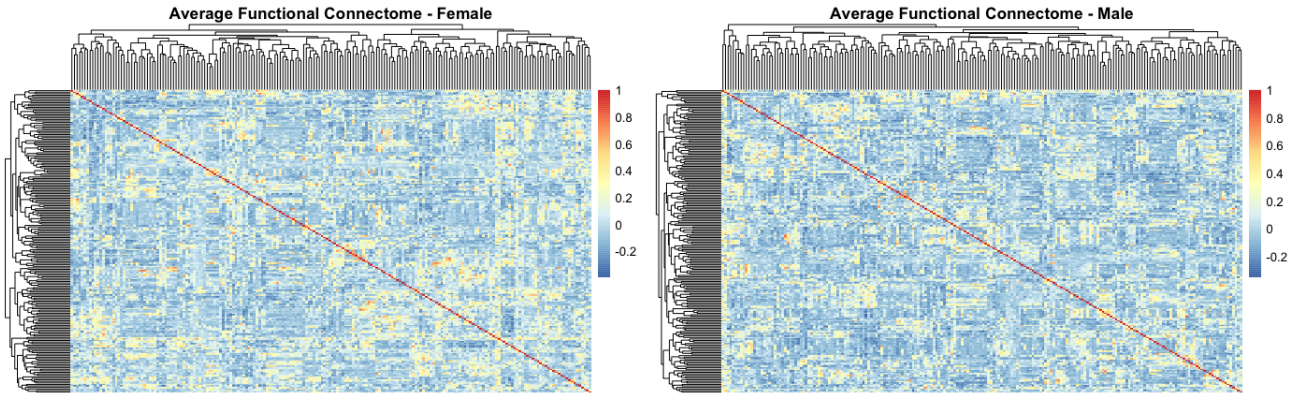


Figure 5: Average Functional Connectome Heatmaps for Males and Females

At the participant-level analysis, the participants' average correlation values show a right-skewed distribution (see Figure 6), with some individuals exhibiting a high average correlation of functional connectivity. The influence of outliers on the bias and variance of the model is taken into consideration during the design process.

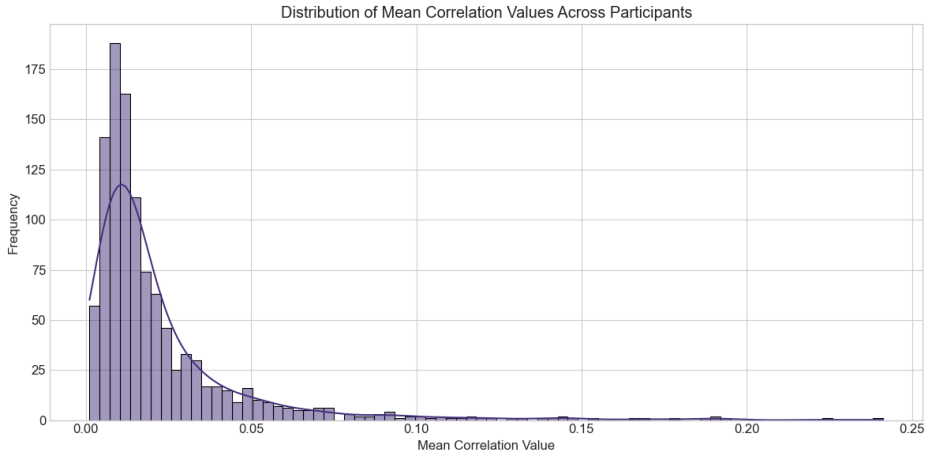


Figure 6: Distribution of Mean Correlation Value

A correlation matrix (Figure 7) was constructed on demographic variables of the training dataset, showing moderate correlation (0.51) between Body Mass Index (BMI) and age, as well as some mild correlation between covariates such 0.27 between externalizing factor scores (reflect dysregulations in rule-breaking behavior) and p-factor scores (reflect the generalized level of psychological difficulties).

These mild correlations among covariates were taken into consideration during modelling to avoid multicollinearity.

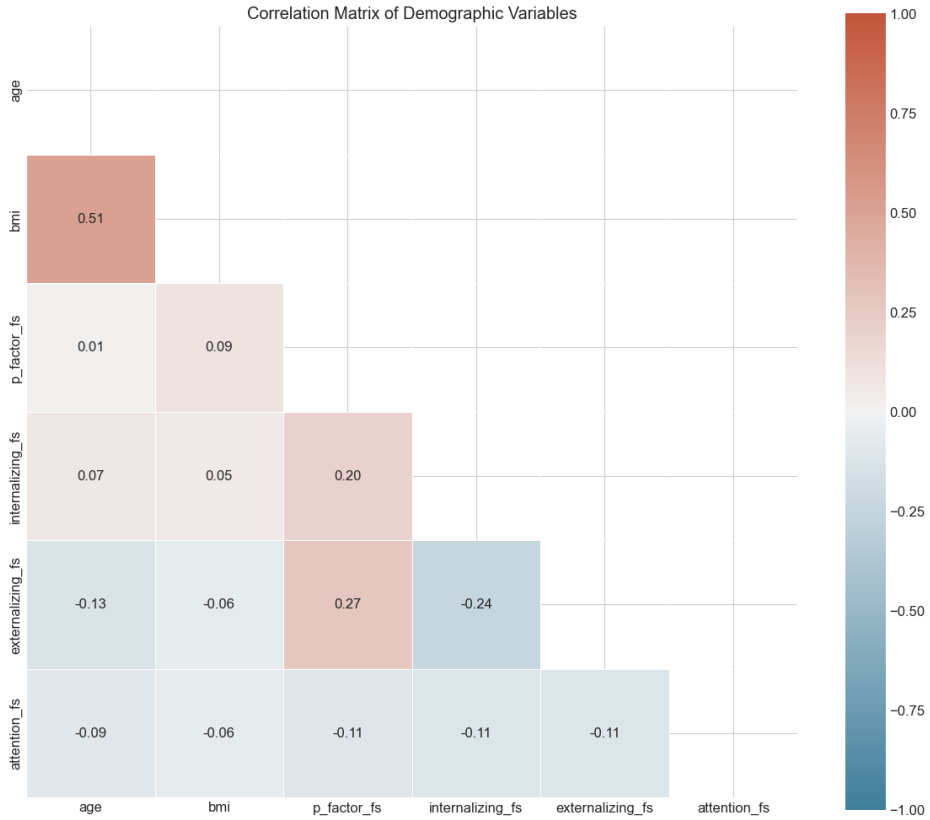


Figure 7: Correlation Matrix of Scores

4 Machine Learning Model and Results

The training dataset is randomly split into 80% training and 20% validation for all model designs. For the baseline reference, no feature engineering was applied, and two linear regression models were fitted (Model 1 - with just the functional connectome matrix as predictors and Model 2 - with the matrix and also the demographic variables). For this model, the BMI missing values were replaced with 0. The missing categorical demographic variables were listed as "not recorded" and removed from the model fit. All the results were tested using 5-fold cross validation on the training data. The mean testing R^2 and Mean Square Error (MSE) for Model 1 is 0.5758 and 4.438, and for Model 2 is 0.5798 and 4.3988, respectively. Both models seems to perform well, but the training R^2 is almost

close to 1, showing serious overfitting problem, indicating that linear model may not be a good option for prediction and miss capturing some nonlinear pattern.

Based on the overfitting issue of linear regression, we suspected that there are too many predictors than observations in the dataset, so we decided to use two dimensionality reduction techniques: PCA and Autoencoder(will be talked in Appendix later). For Lasso Regression, the number of principal components selected was 840, representing a linear combination of 19912 variables, which is needed to explain 95% variance. Also, Lasso model is hypertuned using 5-fold cross validation, and the optimal alpha is 0.2947, which resulted in the best model performance with the testing R^2 of 0.6035 and MSE of 3.9752 (evaluation plot in Appendix).

Other models after PCA were also considered, such as Ridge, Elastic Net model. Ensemble tree models such as Random Forest and XGBoost were also experimented with (interpretation of XGBoost in Appendix). For these models, PCA was performed only on the functional connectome matrix (19900 variables before PCA). A linear PCA resulting in 834 principal components were performed before each model fitting. The imputation method used for these models was to replace the missing BMI with the median and to drop the rows that had missing categorical demographic variables. Similar to previous Lasso, all models were hypertuned using Grid search and 5-fold cross validation. In general, as shown in the results summary in table 1, the elastic net generally got better performance than ensemble methods, but it still got worse results than Lasso. To further improve our performance, we started trying to fit deep learning neural network on the dataset.

Table 1: Comparison of Machine Learning Model Performance

Model	Test MSE	Test R^2
Lasso	**3.9752**	**0.6025**
Elastic Net	5.0485	0.4900
XGBoost	5.5832	0.4354
Random Forest	8.1106	0.3904
Ridge	8.2485	0.3194
Linear Regression (baseline)	4.4382	0.5758

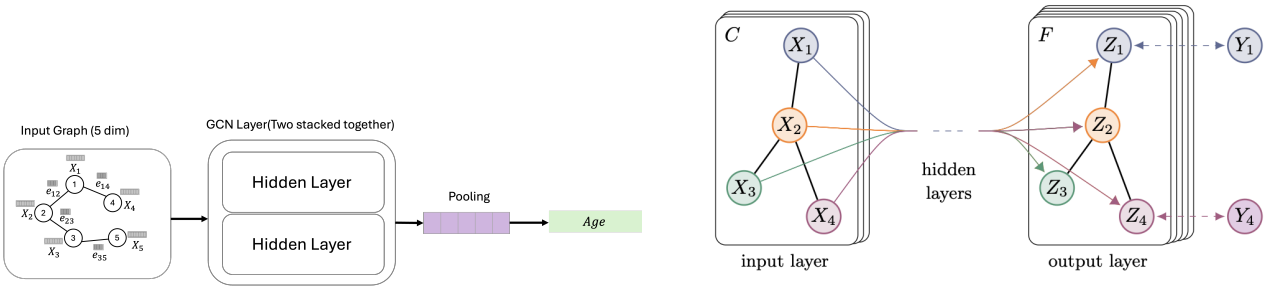
5 Deep Learning Neural Network

In this final project, our team tried several deep learning techniques, including graphical convolutional neural network(GCN)(Kipf and Welling 2016), graphical attention neural network (GAT)(Veličković et al. 2018), multilayer perceptron (MLP) and autoencoder. We will introduce each of those algorithms in the following sections.

5.1 How to Prepare Data for a Graphical Neural Network

Since GNN cares more about the graphical structure of the correlation matrix, we excluded metadata when feeding data into the neural network. Initially, the raw correlation vector, representing the upper triangle of the connectivity matrix (19900 connections between 200 brain regions), undergoes preprocessing: values are clipped to prevent infinities, Fisher z-transformed to stabilize variance, and then standardized (mean-centered and scaled to unit variance). Following this, the full symmetric 200×200 adjacency matrix is reconstructed from the processed vector. The edges of the graph are determined by thresholding this matrix; Only connections with an absolute standardized correlation that exceeds a predefined threshold are retained. The indices of these connections form the edge index, and their corresponding standardized correlation values become the edge weights. Self-loops are also added to each node. Concurrently, node-level features are computed for each brain region based on its connectivity profile within the thresholded graph: these include total connection strength (sum of absolute weights), degree (number of connections), positive strength (sum of positive weights), and negative strength (sum of absolute negative weights). These features, along with a bias term, are stacked into a node feature matrix x. Finally, the function encapsulates the node features (x), the graph structure (edge index, edge weights), and the participant's age (y) into a PyTorch Geometric Data object, creating a single graph representation for that individual suitable for GNN analysis.

5.2 Graphical Convolutional Neural Network



(a) GCN Code Graphical Illustration (b) GCN Structure Overview

Figure 8: Simple Illustration of How GCN Works

This network takes a whole graph (say, one subject's brain-connectivity graph) and predicts a single value (age) in four steps. First, each node starts with five raw features. Second, the two stacked GCNConv layers let every node mix information with its neighbours twice, so local patterns propagate a few hops away and are distilled into a 64-dimensional "hidden" signature per node. Third, a global-mean pool simply averages all these node signatures to create one compact graph-level fingerprint that captures the subject's overall connectivity profile. Finally, a single fully-connected (Linear) layer reads that fingerprint and outputs one number; squeezing it drops the extra dimension so we get a clean scalar age prediction. In short, the network learns neighbour-aware node representations, condenses them into a graph summary, and maps that to the target value.

5.3 Graph Attention Network

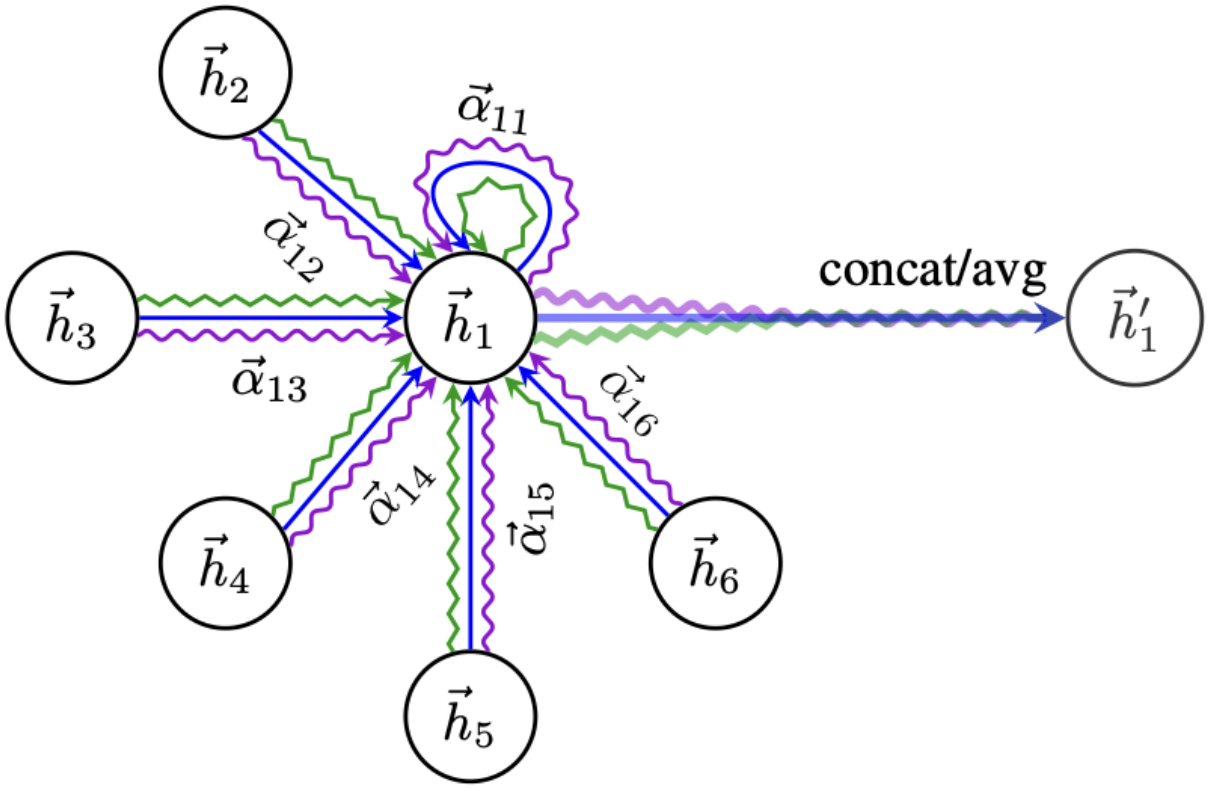


Figure 9: GAT Graphical Illustration

Our graph attention network (GAT) predicts the age of each subject directly from their graph-structured data in four intuitive stages. First, every node enters the model with a five-dimensional feature vector that captures its raw attributes. Second, the graph passes through two successive attention convolution layers. In the first layer four independent "attention heads" learn to weigh each neighbour's contribution differently and then concatenate their findings, giving every node a rich, multi-perspective 32-dimensional representation. The second layer distils those concatenated signals through one more head, producing a streamlined 32-dimensional embedding per node. Third, we use a global-mean pool to compress all node embeddings within a graph into a single graph-level fingerprint that summarises the subject's overall connectivity pattern. Finally, a simple linear re-

gression head maps that fingerprint to one scalar—the predicted age. In essence, as shown in figure 2, each neighbour is presented to h_1 in three different 'perspectives,' and h_1 learns how much to listen to each perspective before blending them into its new embedding. After the new embedding is created, they were aggregated to make prediction.

5.4 How to Prepare Data for Autoencoder

Since the original dataset got 19900 columns for each participant, we convert those 19900 correlation values back to the 200×200 matrix, and now the training dataset should become a $1104 \times 200 \times 200$ matrix. After that, we can fit this matrix into autoencoder to reduce the dimension of the dataset.

5.5 Autoencoder

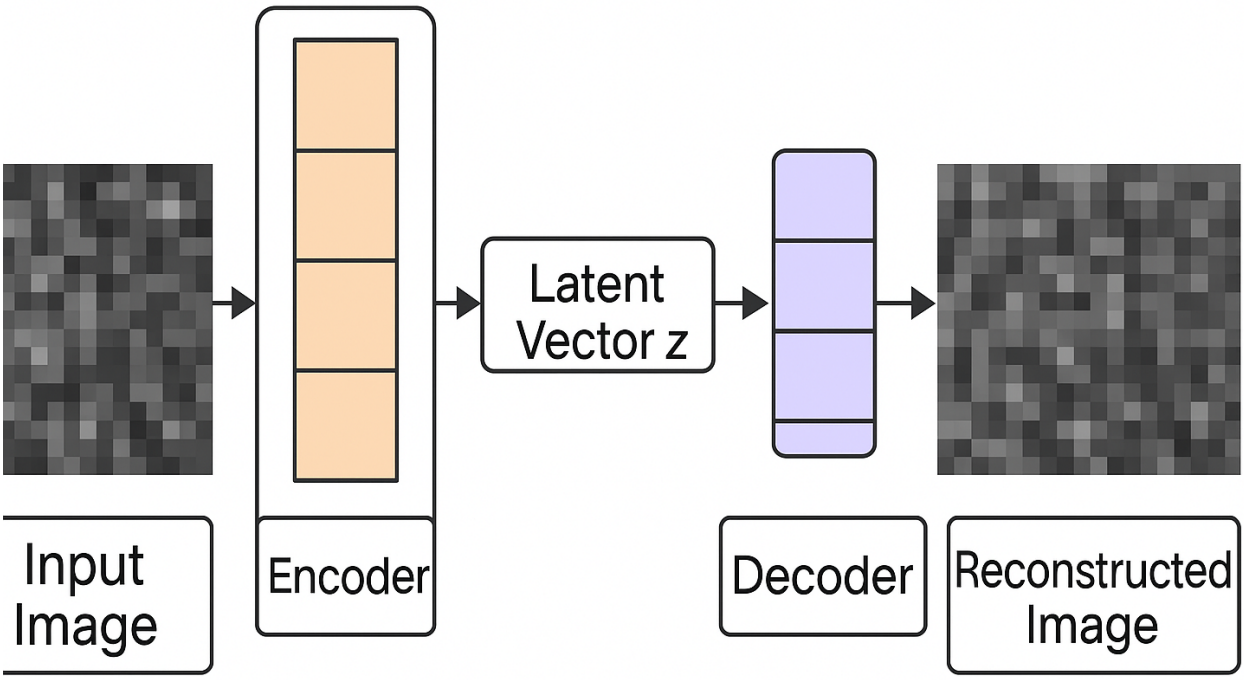


Figure 10: Autoencoder Graphical Illustration

5.5.1 Encoder — squeezing the picture

In our case, single-channel input image (e.g., a 200×200 correlation map) passes through four convolution-plus-ReLU blocks, each halving the spatial size and expanding the channel depth. After the last conv, the data will be converted to a stack of 256 tiny feature maps, 13×13 each. These maps are flattened and fed into a fully connected layer that distils everything down to a 64-number latent vector z (the network's compact snapshot of the image).

5.5.2 Decoder — rebuilding from the snapshot

Another fully connected layer inflates the 64 numbers back to $256 \times 13 \times 13$, reshaping them into a feature stack. Four layers of transposed convolution (a.k.a. deconvolution) then up-sample step by step, reversing the encoder reductions until we recover an image of the original size. The final layer outputs one channel, giving a reconstruction that should look as close as possible to the input. During training, the model tweaks its weights so the reconstructed image minimises a loss (typically mean-squared error) against the original. When it succeeds, the latent vector z becomes a tidy, information-dense representation that can plug into downstream models like Lasso, XGBoost or MLP, while the decoder proves that z really does capture the essentials. For this project, we tried two different autoencoder approaches for the metadata:

1. The first way is to add them as extra dimension when using autoencoder (making it become conditional), then the autoencoder will learn the meta information when reducing the dimension

2. The second way is to use autoencoder only to reduce the dimension of correlation matrix, then after getting the latent vector z, add the extra meta information onto z

5.6 MLP

Based on the reduced output of the autoencoder, we can put the latent vector z in simple MLP. In our code, multilayer perceptron (MLP) predicts age from a

flat feature vector in three quick stages. First, the input passes through a fully connected layer that expands it to 256 hidden units; batch-normalisation steadies the activations and a ReLU adds non-linearity so the network can capture complex patterns. A light 20% dropout then randomly zeros some units, which helps the model generalise instead of memorising the training set. The second hidden layer repeats the process—another 256-unit linear transform followed by ReLU and dropout—allowing the network to refine higher-level feature interactions. Finally, a single-neuron output layer maps the refined 256-dimensional signal to one number, and function *squeeze* removes the surplus dimension so we return a clean scalar age prediction. In short: two stacked $Linear \rightarrow BatchNorm \rightarrow ReLU \rightarrow Dropout$ blocks learn rich, noise-robust representations, and the final linear head translates that representation into the target age.

5.7 Results Summary

Table 2: Comparison of Different Deep Learning Model Performance on Test Data

Model Name	Testing R^2	Testing MSE
Autoencoder + Lasso	**0.540**	**4.611**
Autoencoder + ElasticNet	0.539	4.625
Conditional Autoencoder + Lasso	0.509	4.920
Conditional Autoencoder + ElasticNet	0.514	4.625
MLP with Hypertuning	0.421	6.049

Unfortunately, although GCN and GAT seem to be the most sophisticated option to capture the pattern in the training data, the testing R^2 is negative and MSE is over 10, which gave us even worse results. Based on table 2, fitting Lasso on the reduced dimension using Autoencoder got the best result, which is consistent with our findings in previous machine learning section where Lasso regression on PCA reduced dataset gives the best results. Additionally, using autoencoder as feature engineering did get further improvement for elastic net comparing to PCA + elastic net approach in machine learning section, from 0.49 to 0.539 in R^2.

6 Conclusion

This project embarked on predicting chronological age using resting-state fMRI functional connectome data from the Healthy Brain Network dataset, aiming to identify potential biomarkers of brain development and compare different modeling strategies. Facing the inherent challenge of high-dimensional connectome data ($\approx$ 19,900 features), we systematically explored various approaches, encompassing feature engineering through Principal Component Analysis (PCA) and Autoencoders, alongside a diverse suite of machine learning and deep learning models.

Our investigation compared baseline Linear Regression, regularized methods (Lasso, Ridge, Elastic Net), tree-based ensembles (Random Forest, XGBoost), and advanced deep learning techniques including Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and Multilayer Perceptrons (MLP) coupled with autoencoder-derived features.

The results revealed critical insights into the bias-variance tradeoff inherent in this task. While a simple Linear Regression using raw connectome features achieved a seemingly high test R^2 (≈ 0.58), its near-perfect training R-squared indicated severe overfitting, rendering it unreliable. Exploration into deep learning, particularly GCN and GAT which were theoretically well-suited for graph-structured data, unfortunately yielded poor generalization on the test set (Negative R^2, MSE ¿ 10). Autoencoders provided some benefit, improving Elastic Net performance compared to PCA, but did not lead to the overall best model.

Among all tested configurations, Lasso Regression applied after PCA dimensionality reduction (reducing 19,912 combined connectome and demographic features to 840 components) demonstrated the most robust and effective performance (Test MSE ≈ 3.98, Test $R^2 \approx 0.60$). This approach successfully balanced predictive accuracy with model complexity, mitigating the overfitting observed in the baseline model.

In conclusion, while complex deep learning models require further investigation to successfully apply to this specific dataset, regularized linear models like Lasso, when combined with appropriate dimensionality reduction techniques like PCA that incorporate both connectomic and demographic information, offer a more

reliable and interpretable pathway for predicting age from fMRI data. This work underscores the importance of careful feature engineering and model selection to avoid overfitting in high-dimensional neuroimaging datasets and provides a solid foundation for future research aimed at identifying specific connectomic features driving age prediction.

7 Appendix

7.1 Some Modeling Evaluation Plots

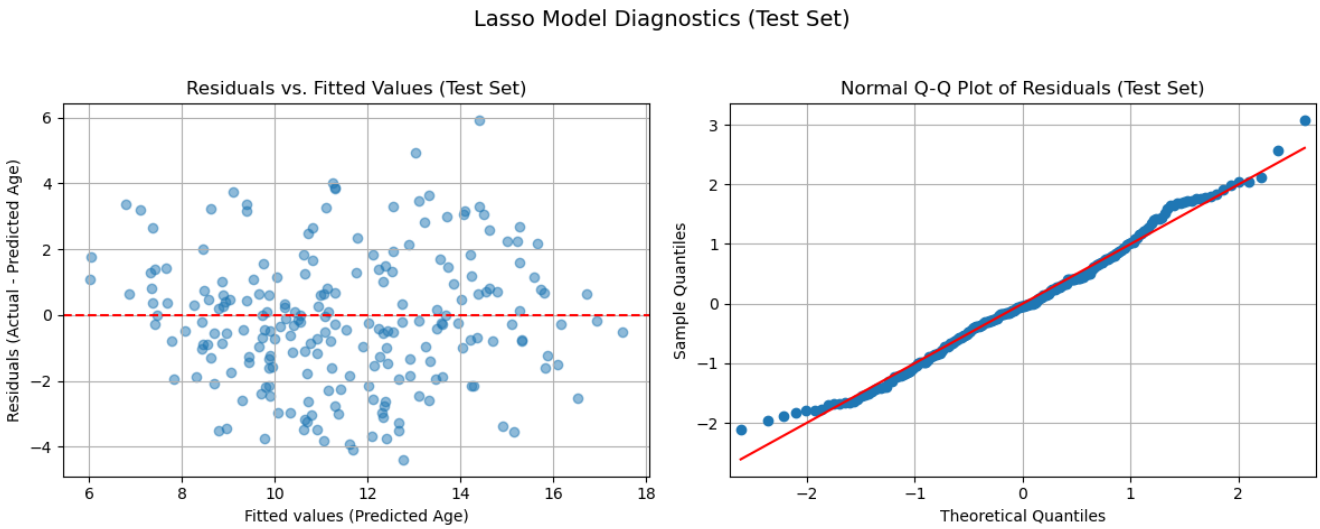


Figure 11: Lasso Model Evaluation

Figure 11 above shows the residual vs fitted plot and QQ plot for the best lasso model. From the residual plot, the residuals are randomly scattered around 0, indicating there is no heteroscedasticity problem, and QQ plot shows almost normal residuals, indicating that the model fitted data well.

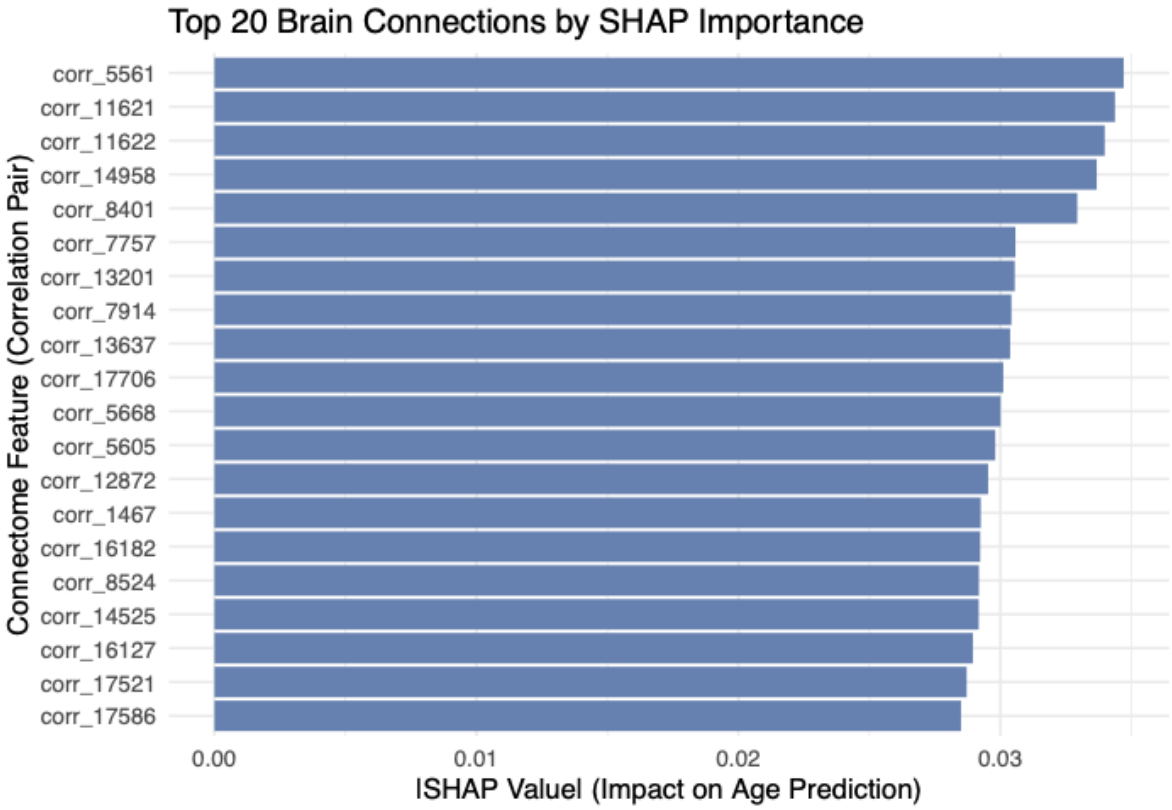


Figure 12: XGBoost Model Interpretation

This Shapley plot shows the 20 features (specific brain connection correlations) that have the largest overall influence on the age prediction model. Taking the absolute value means we were looking at the magnitude of the impact, regardless of whether a high value of that feature increases or decreases the predicted age. A higher absolute SHAP value means the feature has a larger impact on the model's predictions, on average. Based on plot results above, *corr*_5561 seems to have the most impact on the age prediction, followed by *corr*_11621 and *corr*_11622, but since there is no information provided to us about the specific regions, we don't know the correlation is between what two regions, and some more researches are needed to be done on this part.

7.2 Data and Code Availability

All the data used is in this google drive link: drive link to train and test data, and complete code is on Github: link to Github repository

8 References

Kipf, Thomas N. and Max Welling (2016). *Semi-supervised classification with graph convolutional networks*. DOI: 10.48550/arXiv.1609.02907. arXiv: 1609.02907. URL: https://doi.org/10.48550/arXiv.1609.02907.

Menon, Sidharth S. and Karthik Krishnamurthy (2019). "A comparison of static and dynamic functional connectivities for identifying subjects and biological sex using intrinsic individual brain connectivity". In: *Scientific Reports* 9, 5729. DOI: 10.1038/s41598-019-42090-4. URL: https://doi.org/10.1038/s41598-019-42090-4.

Veličković, Petar et al. (2018). *Graph Attention Networks*. DOI: 10.48550/arXiv.1710.10903. arXiv: 1710.10903. URL: https://doi.org/10.48550/arXiv.1710.10903.

WiDS Worldwide, ed. (2025). *Sex-specific differences in the healthy and disordered brain*. URL: https://www.youtube.com/watch?v=KO7YI7j_d-A (visited on 05/05/2025).