**Age Prediction from 2D Functional Brain Networks: WiDS Datathon++ 2025 University Challenge**

Catalin-Ionel Butacu   
*Technical University “Gheorghe Asachi”*Iasi, Romania  
catalin-ionel.butacu@student.tuiasi.ro

Stefana-Beatrice Ghegrhel   
*Technical University “Gheorghe Asachi”*  
 Iasi, Romania   
stefana-beatrice.gherghel@student.tuiasi.ro

***Abstract* — This paper explores predicting age using 2D functional brain networks derived from fMRI data provided by the Healthy Brain Network (HBN). Using brain connectivity matrices and demographic metadata, we developed a regression model with the Elastic Net technique, achieving a root mean square error (RMSE) of 1.81073 on the Kaggle test set. Our results indicate the utility of combining brain connectivity features and metadata for predicting age and underscore the potential of the Elastic Net model in handling high-dimensional data with multicollinearity. Furthermore, this study demonstrates the potential of advanced preprocessing techniques and hyperparameter tuning to achieve state-of-the-art performance in regression tasks related to neuroscience.**

***Keywords— Functional Brain Networks, Elastic Net Regression, Hyperparameter Tuning, fMRI Data Analysis***

# Introduction

Understanding brain development is a cornerstone of neuroscience, with profound implications for the early diagnosis and treatment of neurodevelopmental disorders. The human brain undergoes significant structural and functional changes throughout childhood and adolescence, and these changes are often reflected in the dynamic interactions between different brain regions. Functional brain networks, derived from functional Magnetic Resonance Imaging (fMRI) data, provide a powerful tool for studying these interactions. These networks capture the correlations in neural activity between various brain regions, offering insights into the underlying mechanisms of brain development and function.

The ability to predict age based on brain connectivity data is a challenging machine learning task and a clinically relevant one. Accurate age prediction models can help identify atypical developmental trajectories, which may be indicative of neuropsychiatric disorders such as autism, ADHD, anxiety, and depression. Furthermore, understanding how brain development differs between males and females can provide valuable insights into the sex-specific manifestations of these disorders. This knowledge is crucial for developing targeted interventions and personalized treatment plans.

The Healthy Brain Network (HBN) provides the dataset used in this project, which includes fMRI-derived brain connectivity matrices and demographic information for over 1,500 participants aged 5 to 21 years. The connectivity matrices represent pairwise correlations between 200 brain regions, resulting in a high-dimensional feature space with 19,900 unique features. Additionally, the metadata includes variables such as age, sex, BMI, ethnicity, race, handedness, and psychological scores, which provide complementary information for the prediction task.

# Problem Formulation

The primary objective of this study is to leverage machine learning techniques to build a robust regression model capable of predicting age from functional brain networks and metadata. The problem can be formally defined as follows.

Given a set of functional brain connectivity matrices X and corresponding metadata M, the goal is to predict the age y of each participant. The brain connectivity matrices X are 200x200 symmetric matrices, where each element Xij represents the correlation between the neural activity of brain regions i and j. These matrices are flattened into long-format rows, resulting in 19,900 unique features. The metadata M includes demographic and clinical variables such as sex, BMI, ethnicity, and psychological scores, which are used to complement the brain connectivity data.

One of the key challenges in this project is handling the high dimensionality and multicollinearity inherent in the brain connectivity data. The large number of features (19,900) relative to the number of samples (1,104 training samples) increases the risk of overfitting, while the high correlation between features (multicollinearity) can distort the model's ability to identify the most relevant predictors. To address these challenges, we employ regularization techniques such as Elastic Net, which combines L1 (Lasso) and L2 (Ridge) regularization to manage feature selection and multicollinearity effectively. We also explore other machine learning algorithms, including Bayesian Regression, LASSO, XGBoost, Random Forest, and Support Vector Regression (SVR), to compare their performance in this context.

The motivation behind this work is twofold. First, we aim to contribute to the growing body of research on brain development by developing a predictive model that can accurately estimate age based on functional brain networks. Second, we seek to explore whether there are sex-specific differences in the factors that influence age prediction, which could shed light on the distinct developmental patterns and risk factors for neuropsychiatric disorders in males and females.

By advancing age prediction techniques, this research has the potential to improve early intervention strategies for mental health disorders and provide a deeper understanding of the biological and environmental factors that influence brain development. The findings from this study could also inform the development of personalized treatment plans tailored to the unique developmental trajectories of individuals.

# Data Overview

## Dataset Size and Structure

The dataset used in this project is provided by the Healthy Brain Network (HBN), a large-scale initiative aimed at understanding brain development and mental health in children and adolescents. The dataset is divided into two main subsets:

* Training Samples: 1,104 participants.
* Test Samples: 474 participants.

where each participant’s data includes:

1. Functional Connectivity Matrix: A 200x200 symmetric matrix representing pairwise correlations between 200 brain regions. These matrices are derived from resting-state fMRI data, which measures Blood Oxygen Level-Dependent (BOLD) signals to infer neural activity.
2. Metadata: Demographic and clinical information, including age, sex, BMI, ethnicity, race, handedness, study site, and psychological scores (internalizing, externalizing, and attention scores).

The connectivity matrices are flattened into long-format rows, resulting in 19,900 unique features per participant. This high-dimensional feature space captures the complex relationships between brain regions, providing a rich source of information for age prediction.

## Connectivity Matrix: What It Represents

The functional connectivity matrix is a key component of the dataset. It represents the correlation between the neural activity of different brain regions during a resting state.

1. *Structure of the Matrix*

The matrix structure is a symmetric one, which means the correlation between region i and region j is identical to the correlation between region j and region i. Along the diagonal of the matrix, the values represent the correlation of a brain region with itself, which is always 1, indicating a perfect correlation. The off-diagonal elements, on the other hand, capture the pairwise correlations between different brain regions. These values range from -1 to 1, where a value close to 1 signifies a strong positive correlation, a value close to -1 indicates a strong negative correlation and a value near 0 suggests little to no correlation.

1. *Interpretation of Connectivity Values*

In terms of connectivity interpretation, positive correlations suggest that two brain regions exhibit co-activation patterns, a phenomenon frequently observed among regions within the same functional network, such as the default mode network. Conversely, negative correlations imply an antagonistic relationship, where two regions demonstrate opposing activation patterns, often characteristic of regions belonging to distinct functional networks. Lastly, correlations close to zero indicate a lack of a significant functional relationship between the activity of the two regions.

* 1. *Visualization of the Connectivity Matrix*

To better understand the connectivity patterns, we visualized the connectivity matrix and the distribution of connectivity values from one of our subjects. Below are the key visualizations:

1. *Connectivity Matrix Heatmap*

The connectivity matrix is visualized as a heatmap, where the intensity of the color represents the strength of the correlation. The heatmap provides a clear representation of the pairwise correlations between brain regions.

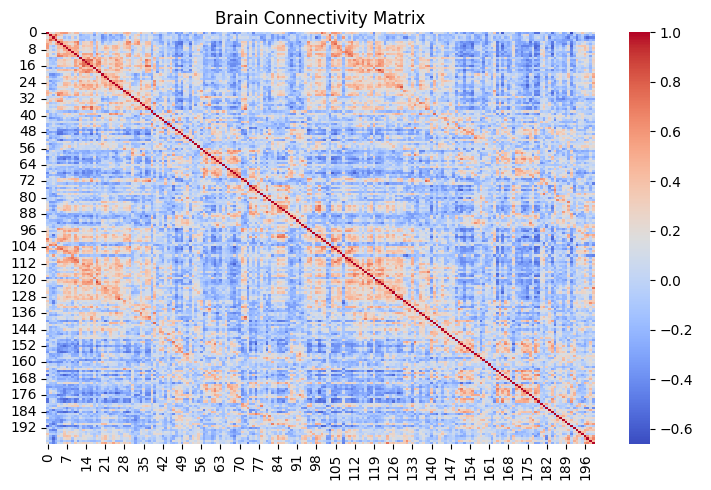


Fig.1 Brain Connectivity Matrix

Figure 1 shows an example of a connectivity matrix heatmap. The diagonal elements are bright (indicating a correlation of 1), while the off-diagonal elements show varying degrees of correlation. Clusters of high correlation (bright regions) indicate strong functional connections between specific brain regions. The color intensity represents the strength of the correlation between brain regions.

1. *Distribution of Connectivity Values*

The distribution of connectivity values was plotted as a histogram to analyze the frequency of different correlation strengths. This visualization helps identify the overall pattern of connectivity in the brain.

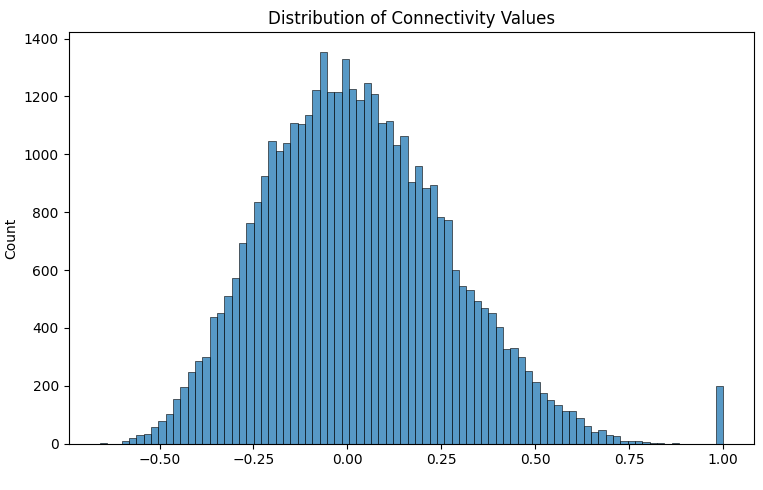


Fig.2 Distribution of Connectivity Values

Figure 2 shows the distribution of connectivity values. The histogram reveals that most correlations are close to zero, with a few strong positive or negative correlations. This indicates that while most brain regions are weakly connected, there are specific regions with strong functional connections. The histogram shows that most correlations are close to zero, with a few strong positive or negative correlations.

# Data Preprocessing

Data preprocessing is a critical step in any machine learning pipeline, especially when dealing with complex datasets such as functional brain connectivity matrices. The goal of our preprocessing data scheme is to transform raw data into a format that is suitable for modeling, ensuring that the data is clean, consistent, and ready for analysis. In this section, we will discuss the key steps taken to preprocess the dataset, including handling missing values, managing outliers, feature exclusion, and encoding categorical variables.

## Features Used

## Brain Connectivity Features:

## The connectivity matrices capture pairwise correlations between 200 brain regions, resulting in 19,900 unique features after flattening the matrices into long-format rows. These features encode the functional connectivity between different regions of the brain, providing rich information about neural activity patterns.

## Metadata Features

## The metadata includes demographic and clinical variables such as age, sex, BMI, ethnicity, race, handedness, study site, and psychological scores (internalizing, externalizing, and attention scores). These features complement the brain connectivity data, capturing individual-level variability relevant to age prediction.

## *Handling Missing Values*

### *Numerical Features:*

For Numerical Features, we employed two imputation techniques to handle missing values: *Mean Imputation* and *KNN Imputation*.

Missing values extracted from fMRI data were replaced using Mean Imputation, where the mean of the respective column was used to fill in the gaps. This method is simple and effective for ensuring the dataset remains complete without introducing significant bias.

For psychological scores—specifically the p\_factor\_fs (general factor linking mental health disorders), internalizing\_fs (emotional or mood problems), externalizing\_fs (behavioral dysregulation), and attention\_fs (inattention and hyperactivity)—we used KNN Imputation.

This technique estimates missing values based on the values of the nearest neighbors, making it more robust and capable of capturing complex relationships between features. KNN Imputation is particularly well-suited for psychological data, as it preserves the nuanced patterns inherent in these scores.

### *Categorical Features:*

For Categorical Features, missing values were replaced with the placeholder "Unknown". This ensures that no information is lost during encoding and allows the model to handle previously unseen categories during inference.

## *Handling Outliers*

Outliers can distort the results of a machine-learning model by introducing noise and bias. In this dataset, outliers were identified as values exceeding three standard deviations from the mean. These extreme values were removed to improve data quality, ensuring that the remaining data (95% of the total) reflected typical patterns. This step reduced the risk of model distortion and improved the robustness of the predictions.

## *Feature Exclusion*

### Not all features in the dataset are equally relevant for the prediction task. Some features were excluded due to a high percentage of missing values or limited relevance to the target variable (age).

### *BMI Exclusion:* Instances lacking BMI data were excluded from the dataset. BMI is an important feature in the prediction task, as it provides information about the physical health of the participants, which may correlate with brain development.

### *Parent Education:* Features related to parent education were excluded due to their high percentage of missing values and limited relevance to predicting age. While parent education may influence cognitive development, its impact on age prediction based on brain connectivity is likely minimal and we’ve preferred to focus only on biological results/analyses.

## *Feature Encoding*

Categorical variables such as sex, ethnicity, and race were one-hot encoded to convert them into numerical formats suitable for regression models. One-hot encoding creates binary columns for each category, allowing the model to interpret categorical information effectively. This transformation ensures that the model can handle categorical data without introducing bias or misinterpretation.

We’ve used the One-Hot Encoding strategy, where each categorical variable was transformed into a set of binary columns, where each column represents a unique category. For example, the "sex" column, which originally contained values "Male" and "Female," was split into two binary columns: "sex\_Male" and "sex\_Female." This encoding prevents the model from assuming any ordinal relationship between categories, which could lead to incorrect interpretations.

# Algorithms

In this study, we evaluated several machine learning algorithms to predict age based on functional brain connectivity matrices and demographic metadata. The primary goal was to identify the most effective model for this high-dimensional regression task. The performance of each model was evaluated using the Root Mean Square Error (RMSE) metric on the Kaggle test set. Below, we discuss the algorithms tested, their configurations, and their respective performance. The results are summarized in Table 1 and will be presented in the next section.

## *Elastic Net Regression*

The Elastic Net model emerged as the best-performing algorithm due to its ability to combine L1 (Lasso) and L2 (Ridge) regularization. This combination allows the model to manage multicollinearity and high-dimensional data effectively, which were key challenges in this project. Implementation involved using scikit-learn’s ElasticNet module, where the alpha parameter controlled the overall regularization strength, and l1\_ratio dictated the balance between L1 and L2 penalties. Hyperparameter tuning was conducted with GridSearchCV over a range of values for these parameters, achieving optimal performance.

## *Bayesian Regression*

Bayesian Regression incorporates prior beliefs about the data distribution and updates these beliefs based on the observed data. This probabilistic approach was implemented using the BayesianRidge module in scikit-learn. The algorithm’s strength lies in its ability to model uncertainty, which is particularly useful for datasets with inherent noise. Despite achieving an RMSE of 1.83686, the computational complexity of Bayesian methods and their reliance on prior assumptions limited their scalability compared to Elastic Net.

## *LASSO Regression*

LASSO Regression, implemented via scikit-learn’s Lasso module, applies L1 regularization to encourage sparsity in feature selection. Two configurations were explored: one with 11-fold cross-validation (RMSE: 1.83884) and another with 5-fold cross-validation (RMSE: 1.87295). The alpha parameter controlled the penalty strength, ensuring that only the most relevant features contributed to the prediction. While effective for dimensionality reduction, LASSO struggled to generalize as well as Elastic Net, particularly when feature correlations were significant.

## *XGBoost*

XGBoost, a gradient boosting framework, was implemented using the XGBRegressor module. Known for its superior performance on structured datasets, the algorithm trains decision trees sequentially, with each tree correcting the errors of its predecessor. Despite its robustness, XGBoost achieved a modest RMSE of 1.95789. The high dimensionality of the dataset likely contributed to overfitting, as the model requires extensive tuning to balance bias and variance.

## *Random Forest*

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting. Implemented with scikit-learn’s RandomForestRegressor, the model aggregated predictions from individual trees using bagging. However, it achieved a subpar RMSE of 2.25605, likely due to the algorithm’s tendency to struggle with sparse and high-dimensional datasets where linear relationships dominate.

## *Support Vector Regression (SVR)*

SVR, implemented using scikit-learn’s SVR module, aims to find a hyperplane that maximizes the margin while minimizing prediction errors. The radial basis function (RBF) kernel was used to capture non-linear patterns in the data. Despite its theoretical robustness, the RMSE of 2.33567 highlights its limitations in handling large-scale, high-dimensional datasets without significant computational overhead.

## *Linear Regression*

Linear Regression served as the baseline model, implemented using scikit-learn’s LinearRegression module. Its simplicity and lack of regularization resulted in an RMSE of 2.34491. While suitable for datasets with strong linear relationships, it proved inadequate for the complexity and multicollinearity inherent in this project’s dataset.

## *Stepwise Regression*

Stepwise Regression iteratively selects or removes predictors based on their statistical significance. While this method simplifies the model, its reliance on p-values makes it prone to overfitting and multicollinearity issues. Implemented manually, the approach yielded an RMSE of 2.41402, comparable to Linear Regression but inferior to regularization-based methods.

## *Decision Tree*

Decision Tree regression, implemented using scikit-learn’s DecisionTreeRegressor, recursively splits the data based on feature thresholds to minimize prediction errors. While intuitive and interpretable, the model suffered from severe overfitting, achieving the worst RMSE of 3.22916. Regularization techniques such as pruning were not sufficient to overcome its limitations in this context.

# Summary of Results

The performance of all tested models is summarized in Table I, indicating that the Elastic Net Regression model outperformed all other models, highlighting the importance of regularization techniques in managing high-dimensional datasets. LASSO and Bayesian regression also performed well, demonstrating the efficacy of linear models with appropriate regularization. Non-linear models, such as Random Forest and SVR, struggled, likely due to overfitting or insufficient tuning for this specific dataset.

| **Model** | **RMSE** | **Notes** |
| --- | --- | --- |
| Elastic Net (Best) | 1.81073 | Best-performing model with tuning. |
| Bayesian Regression | 1.83686 | Performed well, slightly worse than Elastic Net. |
| LASSO (k=11) | 1.83884 | Good performance with 11 splits. |
| LASSO (k=5) | 1.87295 | Slightly worse than k=11. |
| XGBoost | 1.95789 | Moderate performance. |
| Random Forest | 2.25605 | Subpar performance. |
| SVR | 2.33567 | Underperformed compared to Elastic Net. |
| Linear Regression | 2.34491 | Baseline performance. |
| Stepwise Regression | 2.41402 | Poor performance. |
| Decision Tree | 3.22916 | Worst model, likely due to overfitting. |

1. Performance of tested models on Kaggle RMSE metric.

These results suggest that regularization-based linear models are particularly well-suited for high-dimensional data, where managing multicollinearity and preventing overfitting are critical. While non-linear models have the potential to capture complex relationships, they may require more extensive tuning or additional data to perform effectively in this context. This underscores the importance of selecting the right modeling approach based on the dataset's characteristics and the specific challenges of the task.

# Hyperparameter Tuning

Hyperparameter tuning played a crucial role in optimizing the performance of the Elastic Net model. We employed three techniques: GridSearchCV, RandomizedSearchCV, and HalvingGridSearchCV.

### *GridSearchCV:* Conducted an exhaustive search over a specified parameter grid. This method consistently yielded the best results, particularly for the Elastic Net model.

### *RandomizedSearchCV:* Efficiently sampled a fixed number of parameter combinations from a specified distribution, providing a balance between exploration and computational cost.

### *HalvingGridSearchCV:* A newer technique that eliminates poor-performing parameter combinations early in the search process, though it did not outperform GridSearchCV in this project.

# Best Regression Model-ElasticNet

## *Model Overview*

The Elastic Net model combines L1 (Lasso) and L2 (Ridge) regularization, offering a flexible approach to prevent overfitting and manage multicollinearity. This dual regularization approach is particularly beneficial for datasets with a large number of correlated features, such as brain connectivity matrices.

## *Hyperparameter Tuning*

### *Alpha:* Controls the overall strength of regularization. Values were tested logarithmically from 0.001 to 10 to identify the optimal balance between bias and variance.

### *L1 Ratio:* Determines the mix between L1 and L2 regularization. A range of values from 0.1 (mostly L2) to 1 (pure Lasso) was tested.

### *Cross-Validation:* To ensure generalization to unseen data, we performed 5 to 15-fold cross-validation. The best performance was observed with 11 folds, balancing training stability and computational efficiency.

## *Results*

Using GridSearchCV, the Elastic Net model achieved an RMSE of 1.81073 on the Kaggle test set, making it the most accurate model for this task. Its robustness in handling high-dimensional and noisy data underscores its suitability for age prediction based on functional brain networks.

# Conclusion and Future Work

The Elastic Net model demonstrated superior performance in age prediction using 2D functional brain networks and metadata. This success underscores the importance of rigorous preprocessing, feature engineering, and hyperparameter tuning in achieving state-of-the-art results. Future work could explore ensemble methods, such as stacking or blending multiple models, to leverage complementary strengths. Additionally, incorporating other data modalities, such as structural brain networks or genetic information, may enhance predictive accuracy further. By advancing age prediction techniques, this research contributes to a deeper understanding of brain development and its clinical implications.

Future research could focus on interpretability and feature importance analysis to better understand which brain regions or connectivity patterns are most influential in age prediction. Techniques such as Principal Component Analysis (PCA) or SHAP (SHapley Additive exPlanations) could be employed to identify key features and provide insights into the biological mechanisms underlying brain development. This would not only improve the model's transparency but also help clinicians and researchers identify potential biomarkers for neurodevelopmental disorders. By combining predictive accuracy with interpretability, this line of research could bridge the gap between machine learning and clinical applications, paving the way for more targeted and effective treatments.

# 

# References

[1] Scikit-learn: Machine Learning in Python. [Online]. Available:

[https://scikit-learn.org](https://scikit-learn.org/)/

[2] M. Gabrielscu, "Fundamentals of Machine Learning",Lecture Notes

[3] F. Leon, "Artificial Intelligence," Lecture Notes

[4] "Sklearn Regression Models," JavaTpoint.

[www.javatpoint.com/sklearn-regression-models](http://www.javatpoint.com/sklearn-regression-models)

[5] "WiDS Datathon 2025 University Competition Dataset," Kaggle.

[www.kaggle.com/competitions/widsdatathon2025-university/data](https://www.kaggle.com/competitions/widsdatathon2025-university/data)

[6] "WBHI Resources," University of California, Santa Barbara.

<https://wbhi.ucsb.edu/resources>

[7] "TXYZ AI Platform," for research and analyse of paper documents

<https://app.txyz.ai/>