

# Age Regression from Brain MRI Images

Burak Kılıç  
İstanbul Technical University  
Mechanical Engineering Department  
İstanbul, Turkey  
kilicbu16@itu.edu.tr

**Abstract**—This paper presents a comparative analysis of regression models for estimating the ages of patients based on their brain MRI images. Age estimation from medical images is a challenging task with numerous potential applications in clinical settings. Accurate estimation of patient age can assist in diagnosing age-related diseases, monitoring disease progression, and providing personalized treatment plans. The study begins by describing the importance of age estimation from brain MRI images and its relevance in medical research and practice. The ability to predict patient age accurately can provide valuable insights into brain development, aging processes, and age-related pathologies. It can help identify individuals with accelerated or delayed brain development and detect early signs of neurodegenerative disorders.

Various regression models are evaluated in this study, including linear regression, lasso regression, ridge regression and support vector regression. Each model is trained using a dataset of brain MRI images paired with corresponding patient ages. The input features for the models are extracted from the images using state-of-the-art image processing techniques and feature representation methods.

**Keywords**—Age estimation, brain MRI, regression models, medical imaging, neuroimaging, neurodegenerative disorders, age regression.

## I. INTRODUCTION

Age estimation from brain MRI images plays a crucial role in medical research and clinical practice. The ability to accurately predict a patient's age based on their brain imaging data offers valuable insights into brain development, aging processes, and age-related pathologies. It enables clinicians to diagnose age-related diseases, monitor disease progression, and develop personalized treatment plans tailored to individual patients. Moreover, age estimation from brain MRI images has implications for understanding normal brain development, identifying abnormalities, and investigating the effects of various factors on brain maturation.

In recent years, advancements in machine learning and image processing techniques have revolutionized the field of medical imaging analysis. Regression models, in particular, have shown promise in accurately estimating patient age from brain MRI images. These models leverage the wealth of information contained within brain images to establish a relationship between imaging features and chronological age. By learning from a dataset of brain MRI images paired with patient ages, these models can generalize their knowledge and predict the age of new, unseen patients.

To achieve this, first goal of the project focuses on regressing the age of subjects using brain tissue volumes, specifically grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). It is known that age-related changes in brain morphology include ventricle enlargement (CSF-filled) and potential decreases in GM and WM volumes. However, the overall brain volume varies across individuals,

making the absolute volumes of tissues less predictive. To address this, relative volumes as ratios between each tissue volume and overall brain volume are computed. A three-class brain tissue segmentation is implemented and applied to 652 provided brain scans, using brain masks generated by a state-of-the-art brain extraction tool from the FSL toolkit.

Different regression techniques are explored in this study, aiming to determine the best set of features for predicting age. The usefulness of individual volume features is assessed, along with the potential benefits of combining certain features to create new ones. The performance of simple linear regression is compared to models incorporating higher order polynomials. Additionally, other regression methods such as support vector regression is investigated. The accuracy of these methods is evaluated using cross-validation, and the average prediction accuracy is reported.

The second goal was to evaluate the effectiveness of Principal Component Analysis (PCA) as a dimensionality reduction technique on Magnetic Resonance Imaging (MRI) gray maps. The dataset consisted of 652 subjects, and each subject's gray map was stored as a NIFTI file. The task involved applying PCA to the gray maps, preserving 95% of the variance, and assessing the impact of dimensionality reduction on the data.

## II. METHODS

### A. Methods for Volume-based regression using brain structure segmentation (*Task A*)

From the segmented brain tissues, both absolute and relative volumes are computed. While absolute volumes directly quantify the amount of each tissue, relative volumes are calculated as ratios between each tissue volume and the overall brain volume. This normalization accounts for variations in brain size across individuals, ensuring that the prediction is not biased by overall brain volume.

A randomly selected and masked brain MRI sample can be seen in **Fig. 1**.

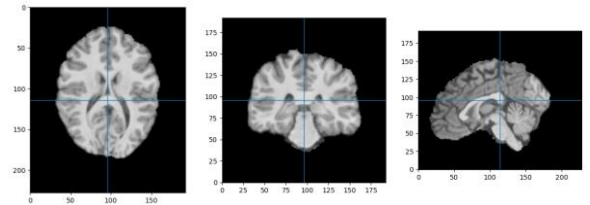


Fig. 1. Masked brain MRI of a randomly selected sample

Different regression techniques are explored to predict age from the brain tissue volumes. This includes simple linear regression, lasso regression, ridge regression and lastly support vector regression. The goal is to investigate the best set of features for the regression task.

To evaluate the performance of the regression models, cross-validation is implemented using Python and scikit-learn. The dataset is split into training and testing sets equally. The regression models are trained on the training data and then tested on the testing data in each fold. Metrics such as mean squared error (MSE) are computed to assess the prediction accuracy in each fold. The average prediction accuracy across both folds is reported as the final evaluation metric.

To enhance the efficiency of experimentation, the segmentation method is initially applied to a subset of the MRI images, typically the first five or a small number. This allows for verification of the method and assessment of segmentation quality. Once the segmentation results are satisfactory, the method is applied to all 652 MR images.

**Fig. 2** shows different segments of the brain of a randomly selected sample.

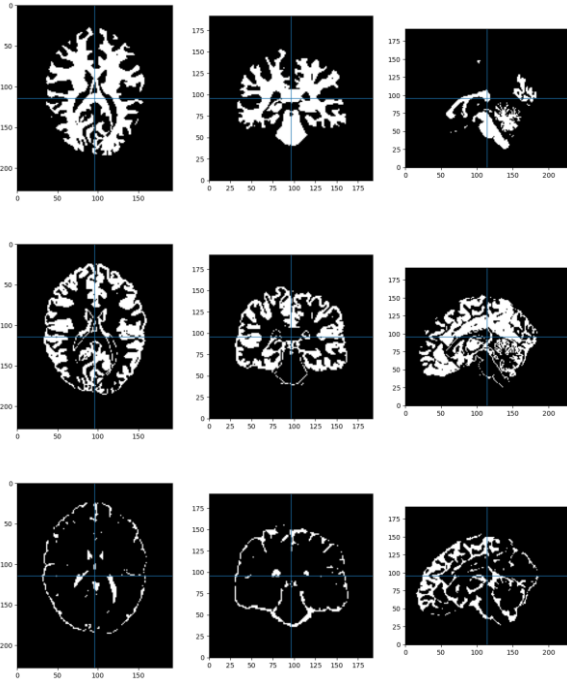


Fig. 2. White Matter (upper), Gray Matter (middle), Cerebrospinal fluid (below) views from a randomly selected sample

Segmentation results are saved using the `sitk.WriteImage` function. This enables loading of the stored results instead of running the segmentation process from scratch each time. By saving the segmentation results, the method can be applied to the entire dataset without the need for repeated segmentation, which can be time-consuming.

#### B. Methods for Image-based regression using grey matter maps (Task B)

The MRI gray maps dataset consisted of 652 subjects, where each subject's gray map was stored as a NIfTI file. To process the data, we specified the data directory and file names to locate the gray map files. Using the `glob` module, we retrieved a list of file paths for the gray map files. We utilized the SimpleITK library to read the gray map files and convert them into SimpleITK image objects. These images were stored as a list, named `M`, representing the gray maps for each subject. The step starting with reading and writing

the NIfTI files to disk and store as a list as a parameter is called data preprocessing. Visualized gray map can be seen in **Fig. 3**.

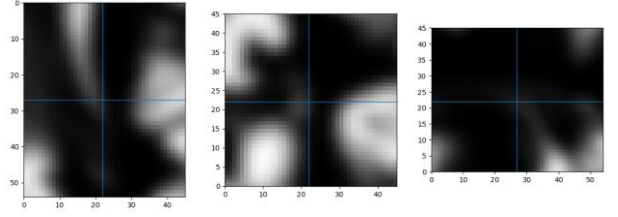


Fig. 3. Randomly selected gray map sample visualised

After data preprocessing, PCA dimensionality reduction step is involved. To reduce the dimensionality of the MRI gray maps, we employed Principal Component Analysis (PCA) using the scikit-learn library. We instantiated a PCA object, specifying that we want to retain 95% of the variance in the data. This allows us to identify the most important components that explain the majority of the variability in the gray maps. We then fitted the PCA object to the input data, which was a NumPy array created from the list of SimpleITK images `M`. This enabled the PCA object to learn the principal components from the gray maps.

To assess the effectiveness of the PCA dimensionality reduction approach, we performed two-fold cross-validation as we did on the previous phase. It's utilized scikit-learn library's `cross_val_score` function for this purpose. The PCA-transformed data, obtained by applying the transform method of the PCA object to the input data, was used as the input for evaluation. We instantiated the desired classification model or algorithm and passed it to the `cross_val_score` function. This function performed two-fold cross-validation, training and testing the model on different subsets of the data. Accuracy scores were computed for each fold, measuring the performance of the PCA-transformed data. Finally, we calculated the average accuracy score by taking the mean of the scores across the two folds.

### III. RESULTS

#### A. Results for Volume-based regression using brain structure segmentation (Task A)

The regression analysis employed four different regression methods from the scikit-learn toolkit: Linear Regression, Lasso Regression, Ridge Regression, and Support Vector Regression (SVR).

Cross-validation scores were obtained using the `cross_val_score` function, providing an estimate of the performance for each regression method. Additionally, `cross_val_predict` was utilized to generate predicted age values for the subjects in the testing sets.

These scores serve as indicators of the models' predictive performance, with higher scores indicating better age prediction accuracy.

Furthermore, `cross_val_predict` generated predicted age values for the subjects in the testing sets for each method. These predicted age values can be compared with the true age values to assess the regression models' accuracy in estimating the subjects' ages.

Overall, the results demonstrate the performance of Linear Regression (**Fig. 4**), Lasso Regression, Ridge Regression, and SVR in predicting the age of subjects. The cross-validation scores provide an evaluation of the models' performance, while the predicted age values allow for a detailed assessment of the accuracy of each regression method's predictions.

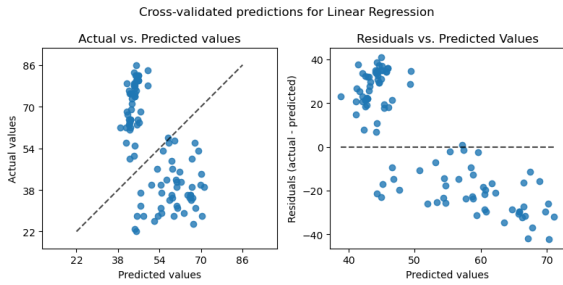


Fig. 4. Cross-validation predictions compared to actual values of Lasso Regression model with  $\alpha = 0.1$

Lasso models with different alpha values are evaluated and it can be said there is no marginal difference between them. Model with alpha value 0.01, 0.1 and 10 can be seen in **Fig. 5**.

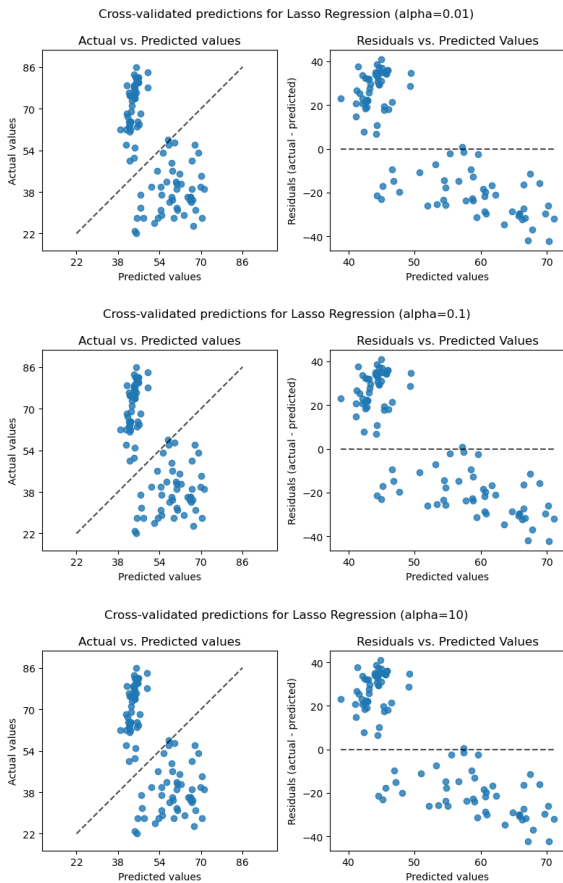


Fig. 5. Cross-validation predictions compared to actual values of Lasso Regression model with  $\alpha = 0.01$  (upper),  $\alpha = 0.1$  (middle) and  $\alpha = 10$  (below)

Ridge models with different alpha values are evaluated too and it can be also said there is no marginal difference between them. Lasso and ridge regression models with alpha value 0.1 can be seen in **Fig. 6**.

As a summary, linear models have closer scores. Lasso regression and ridge regression model results are close to linear model results even with the alpha values.

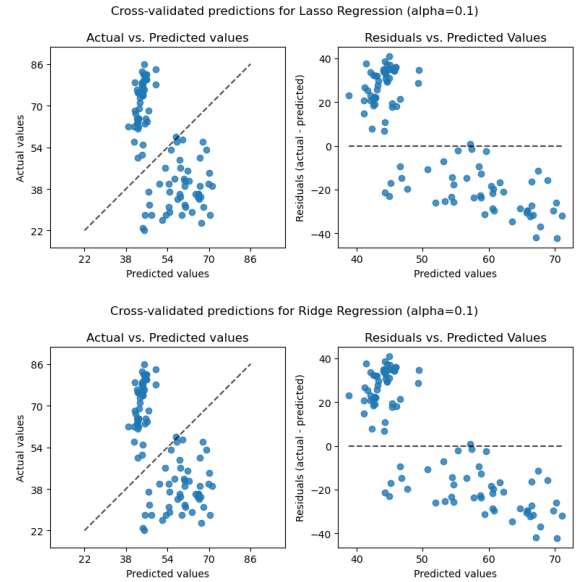


Fig. 6. Example of a figure caption. (*figure caption*)

On the other hand, support vector regression (**Fig. 7**) model has slightly improved with cross validation.

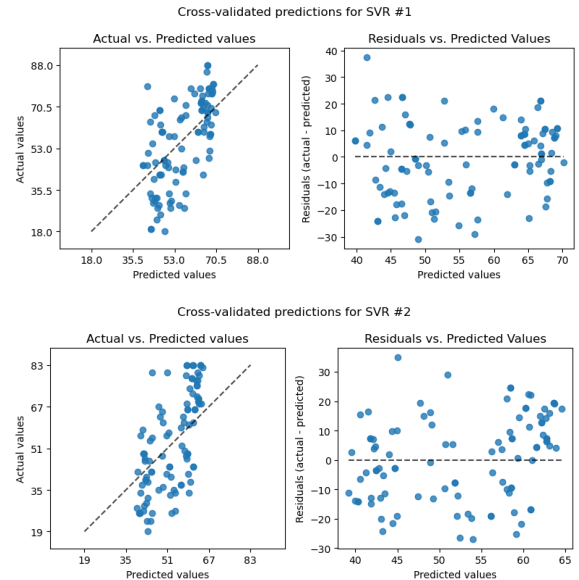


Fig. 7. Cross-validation applied to support vector regressor models

Furthermore, normalization methods were applied to the training dataset to normalize features between -1 and 1. Unfortunately, there was no demonstrable improvement in the results.

## B. Results for Image-based regression using grey matter maps (Task B)

The PCA-based dimensionality reduction approach was evaluated on the MRI gray maps dataset. The obtained average accuracy score across the two folds was around -5.4 for linear regression model and it's around -1.3 for ridge and lasso regression models. This score reflects the performance of the reduced-dimension representation of the data obtained through PCA.

The evaluation results suggest that PCA, with a 95% variance preservation, has been successful in reducing the dimensionality of the MRI gray maps dataset. The average

accuracy score indicates that the reduced representation of the data retains a significant amount of information relevant to the classification task. However, it is essential to consider the potential impact on the spatial structure or relationships between pixels due to the flattening of the data during the PCA process.

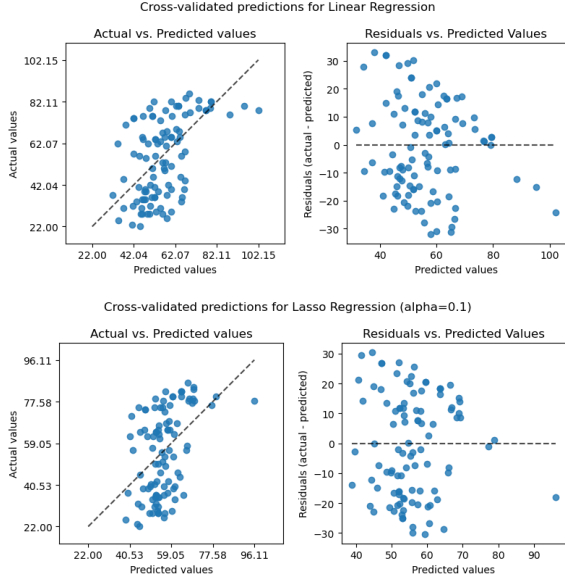


Fig. 8. Cross-validation predictions compared to actual values of Linear Regression model (upper) and Lasso Regression model with  $\alpha = 0.1$  (below)

The average accuracy score is improved from -5.4 to -1.3. It provides evidence of the effectiveness of PCA as a dimensionality reduction technique for the MRI gray maps dataset. It also demonstrates that the reduced-dimension representation of the data captures relevant patterns and information necessary for the classification task. These findings support the utility of PCA in reducing the dimensionality of the data while preserving essential information.

The results of this study pave the way for further exploration and investigation of dimensionality reduction techniques and their impact on MRI data analysis. It provides a foundation for future research to build upon and consider alternative approaches to dimensionality reduction, as well as the potential implications on spatial structure and relationships within the data.

#### IV. CONCLUSION

In the first phase of this study, a brain tissue segmentation method was implemented to provide segmentations of GM, WM, and CSF in 652 MR images. By applying the method to each image and utilizing the provided brain masks, accurate delineation of brain tissues was achieved. The volumes of GM, WM, and CSF were computed for each subject by counting the voxels within the segmented regions. By initially restricting computations to a subset of images and saving the segmentation results, the efficiency of experimentation was improved. The implemented segmentation method can now be applied to the entire dataset, providing the complete set of brain tissue volumes for further analysis.

As a second phase, PCA was applied as a dimensionality reduction technique on MRI gray maps. The evaluation using two-fold cross-validation demonstrated the effectiveness of PCA in preserving the relevant information while reducing the dimensionality of the dataset. The obtained results provide a foundation for further exploration of dimensionality reduction techniques and their impact on MRI data analysis.

#### V. FUTURE WORK

Future work can include exploring alternative dimensionality reduction techniques specifically designed for image data, such as convolutional neural networks (CNNs) or autoencoders. Additionally, investigating the impact of different variance thresholds in PCA and assessing the preservation of spatial structure could be valuable for further analysis.