Abstract #206 | Oral or Poster presentation

Enhancing brain age prediction: stacking models and insights on sample size

L. Backes[1], S. Eickhoff[1, 2], C. Rubert[2], C. Philllips[3], G. Antonopoulos[1], K. Patil[1]

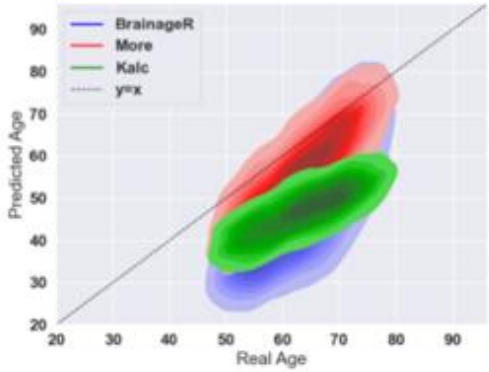
[1] *forschungszentrum Jülich, Institute of Neurosciences and Medicine (INM), Jülich, North Rhine-Westphalia, Germany*
[2] *Heinrich-Heine-University, Institute of Systems Neuroscience, Düsseldorf, North Rhine-Westphalia, Germany*
[3] *University of Liège, GIGA CRC Human Imaging, Liege, Belgium*

As life expectancy increases, the need for reliable biomarkers of biological age becomes crucial. Brain age prediction using MRI offers a promising solution, yet current models struggle with generalization due to site and scanner variability, limiting their clinical utility.

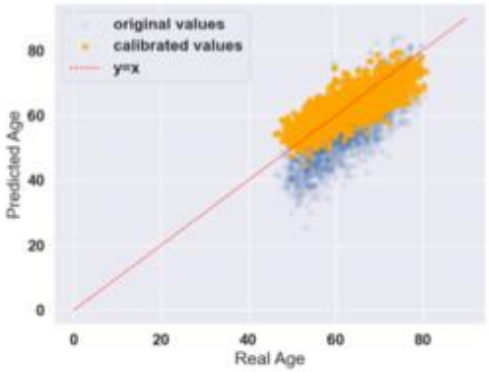
In this study, we evaluated three different Gaussian Process Regression (GPR) models—Kalc, brainageR, and More—trained on the same seven datasets (IXI, AIBL, DLBS, GSP, NKIRSE, OASIS-1, SALD) to predict chronological age from healthy brain scans. These models differ in preprocessing pipelines and feature extraction methods: Kalc uses CAT12 for GM/WM extraction with PCA and ensemble strategies, brainageR relies on SPM12 for segmentation and normalization with PCA (80% variance retained), and More extracts GM volume features from CAT12, smoothing and resampling before applying PCA (100% variance retained). We tested their performance on an out-of-sample UKB dataset (N=6883), performed individual model calibration (N=4818 training, N=2065 test), and applied stacking using Ridge regression and Random Forest, systematically increasing training sample sizes (100 to 4000) to assess performance stability.

Among the three models, More performed best out-of-the-box (MAE=7.35, r=0.81), while Kalc and brainageR showed significantly larger errors. Linear calibration improved all models, with More maintaining superior performance (MAE=3.50, r=0.81). Stacking models further enhanced predictions, with Ridge regression performing well at small sample sizes but plateauing at N=500, whereas Random Forest required larger training sets but eventually outperformed Ridge beyond 1500 samples. Including sex as a feature slightly improved performance but with minimal impact.

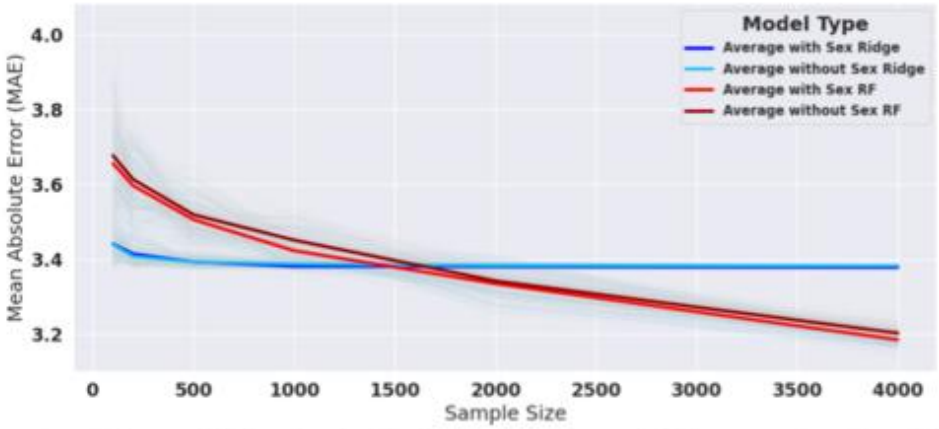
Our findings highlight the variability in pre-trained model performance and emphasize the effectiveness of stacking models, particularly Random Forest, in improving brain age prediction. This approach encourages further integration of meta-learning strategies and sex-specific considerations in brain age research, ultimately enhancing the clinical applicability of these models.



(a) Contour Plot of Predicted Age vs Real Age for the three models.



(b) Scatter Plot of Predicted Age vs Real Age for Calibrated More.



(c) Comparison of the test MAE vs Sample Size for Random Forest and Ridge regression with and without Sex.

Figure 1: Comparison of brain age prediction performance across models and conditions. (a) Contour plots of predicted age versus real age for the three models, highlighting prediction accuracy and distribution. (b) Scatter plots of predicted age versus real age with the original values predicted by More (blue) as well as the calibrated values (orange) obtain with linear regression. (c) Test MAE as a function of sample size for Random Forest and Ridge regression models, comparing performance with and without incorporating sex as a feature.

Comparison of brain age prediction performance across models and

Table 1: MAE, R^2 and pearson correlation for each individual models, linear regression using each model and stacking models (with and without sex)

Model	**MAE**	R^2	**Pearson Correlation**
BrainageR	15.12	-4.40	0.67
Enigma	16.60	-4.42	0.75
More	7.35	-0.40	0.81
linear model (BrainageR)	4.56	0.44	0.66
linear model (Enigma)	4.01	0.55	0.74
linear model (More)	3.50	0.65	0.81
Ridge (with sex)	3.38	0.67	0.82
Ridge (without sex)	3.38	0.67	0.82
RF (with sex)	3.19	0.69	0.83
RF (without sex)	3.20	0.68	0.83

Metrics for each individual models, linear regression using each model and s