Enhancing Brain Age Prediction:
100 is all you need

Backes Lucas[1,2], Eickhoff Simon[1,2], Rubbert Christian[3], Phillips Christophe[4], Antonopoulos Georgios[1,2] & Patil Kaustubh[1,2]

[1]Institute of Systems Neuroscience, Heinrich Heine University Düsseldorf, Düsseldorf, Germany;
[2]Institute of Neuroscience and Medicine (INM-7: Brain and Behaviour), Research Centre Jülich, Jülich, Germany
[3]Department of Diagnostic and Interventional Radiology, Medical Faculty and University Hospital Düsseldorf, Heinrich-Heine-University Düsseldorf, Germany
[4]GIGA CRC Human Imaging, University of Liège, Belgium
l.backes@fz-juelich.de

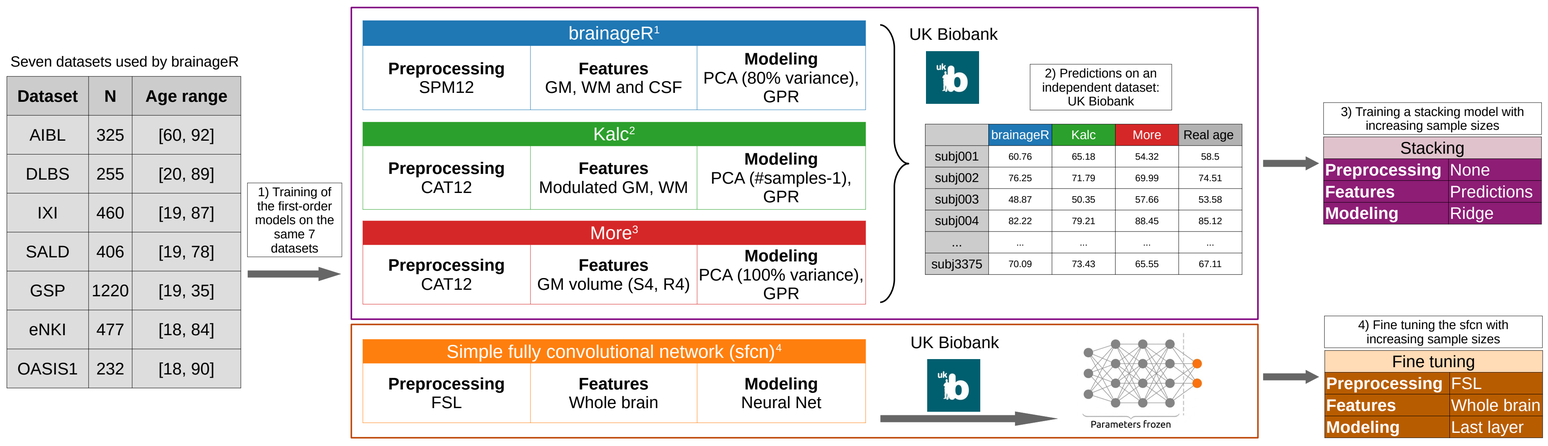
Introduction

- MRI can help us produce a biomarker for an overall 'brain age' such that it more accurately measures disease and mortality risks than chronological age.
- However, due to site/scanner differences as well as demographic differences, brain age models do not usually work well on new datasets that were not used for training. This drawback prevents clinical use of 'brain age'' as an informative and relevant biomarker.
- Stacking and fine-tuning can help us improve generalization.

Methods

- Train different algorithms on the same seven datasets.
- Use the UK Biobank (UKB) as a new test site.
- Evaluate stacking of Machine Learning (ML) models and fine-tuning of a Deep Learning (DL) model using increasing size of UKB training samples.
- Compare the enhanced models with the ouf-of-box models on unseen data of underrepresented population.

Schematic



1) Each base (first-level) model differs in its preprocessing pipeline and in the features it uses. The more diverse these models are, the more effective the stacking ensemble becomes.

2) We observe that each model produces substantially different predictions when evaluated on an unseen dataset such as the UKB.

3) We trained the models using sample sizes ranging from 10 to 150, and then evaluated them on a held-out test set from the UKB.

Results

Out-of-box

1

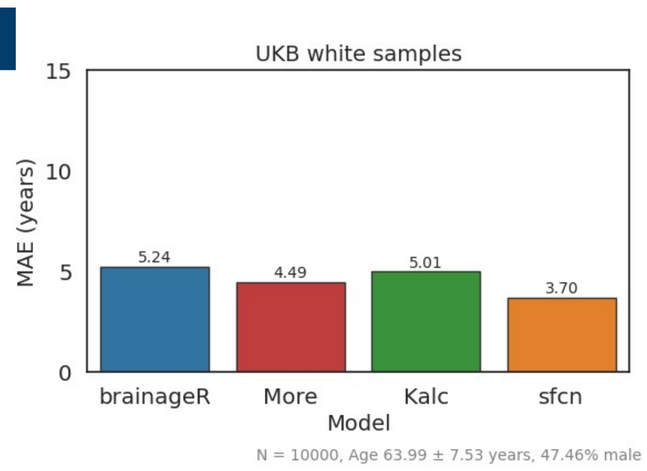


Figure 1: Bar plots of the MAE for all models tested on the white samples of UKB

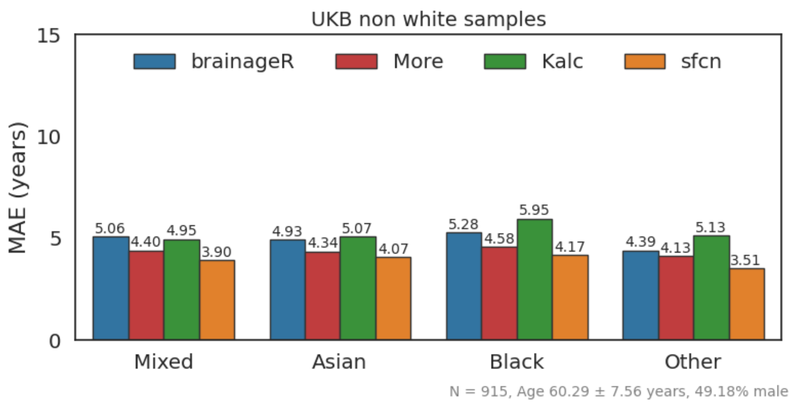


Figure 2: Bar plot of the MAE for all models tested on the non white samples of UKB (Mixed, Asian and Asian british, Black and Black british and Other race)

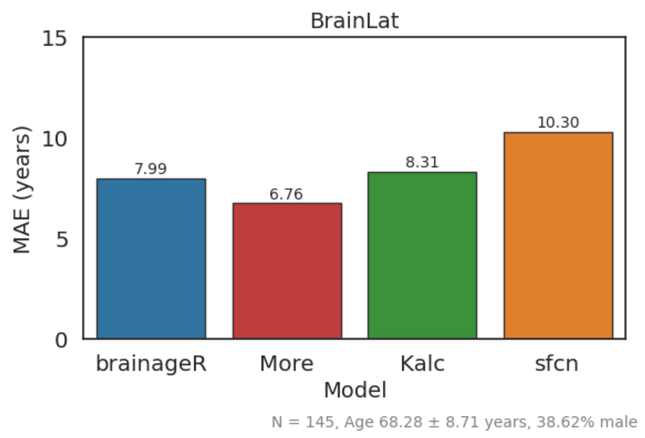


Figure 3: Bar plot of the MAE for all models tested on the latin American dataset BrainLat

Enhanced

2

- Evaluated calibration (single model), stacking (multiple models) and fine tuning on the same 10.000 samples as the out-of-box models.
- Same stratified UKB samples (white healthy individuals) were used for model enhancement (calibration, stacking and fine-tuning).
- For stacking, **Ridge regression** and **Random Forest** were compared. For fine-tuning, we compared **validation** apporach and **full training** approach.
- *Ridge regression* and the validation approach were retained.

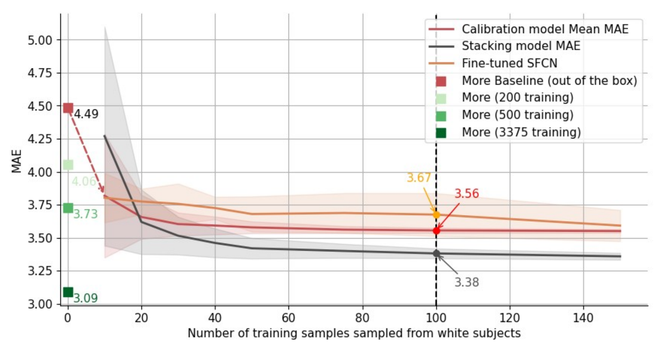


Figure 4: Stacking, calibration and fine-tuning curves of MAE w.r.t the training sample size

- For comparaison, we trained the best single ML model (*More*) from scratch with 200, 500 and 3375 stratified samples from UKB and tested on the same 10.000 samples.
- For fine tuning, we used 80% of the samples as training and the remaining 20% as validation for early stopping. Models at the best epoch were kept.

3

Train → Test	BrainLat	Mixed	Asian	Black	Others	Mean
BrainLat	**5.37**	7.47	6.99	6.47	5.40	6.34
Mixed	6.38	**3.60**	3.41	3.40	3.36	4.03
Asian	5.89	3.61	**3.35**	3.34	3.11	3.86
Black	5.62	3.77	3.48	**3.20**	3.12	3.84
Others	5.65	3.67	3.35	**3.20**	**3.10**	3.79
Mean	5.78	4.43	4.12	3.92	3.62	4.37

Table1: Cross datasets results of stacking models trained on 100 samples from one subgroup and tested on the remaining subrgoups.

- We then trained a new stacking model on 100 samples of each subgroups (non-white subgroups from UKB and BrainLat) and tested on the remaining.
- Each number is the average of 50 different train/test splits.

Discussion

1

- Among the three classical ML models, *More* yielded the best results for the white samples from UKB (MAE=4.49), outperforming *Kalc* (MAE=5.01) and *brainageR* (MAE=5.24).
- *More* also outperformed them when looking and non-white subjects (Figure 2 and 3).
- The *sfcn* model from Peng generalized best as an "out-of-box" model for the UKB, achieving lowest MAE of 3.70 for white samples. However is being surpassed by *More* for BrainLat (Figure 3: *sfcn* MAE=10.30; *More* MAE=6.76).

2

- *Ridge regression* (MAE=3.38) outperformed calibration method (MAE=3.56) and the fine-tuning method (MAE=3.67) at 100 training samples.
- All stacking, calibration and fine tuned models performed better than out-of-box *More* (MAE=4.49*)* **and** *More* trained from scratch with 200 training samples (MAE=4.06) after only 20 training samples.
- At low data regime (< 20 samples), calibration and fine-tuning appear to perform better on average then stacking.
- From 40 samples onward, the stacking method outperformed the other methods, as well as *More* trained with 500 samples.

3

- Knowing that performance did not increase a lot after 100 samples we decided to take 100 training samples of each underrepresented group and tested the performance on 40 held out samples.
- For UKB, It seems like for all subgroups, stacking (even if trained on a different subgroup) outperforms the best out-of-box models. For instance, it is true for the Asian subrgoup (stacking trained on Black: MAE=3.48; *More*: MAE=4.34; *sfcn*: MAE=4.07).
- For BrainLat, even the models trained on the UKB subgroups performed better than the out-of-box models (*More*: MAE=6.76 ; stacking trained on Black: MAE=5.62).

References: [1]Sudlow, Cathie et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." PLoS medicine vol. 12,3 e1001779. 31 Mar. 2015, doi:10.1371/journal.pmed.1001779; [2]Kalc, Polona et al. "BrainAGE: Revisited and reframed machine learning workflow." Human brain mapping vol. 45,3 (2024): e26632. doi:10.1002/hbm.26632; [3]More, Shammi et al. Brain-age prediction: A systematic comparison of machine learning workflows, NeuroImage, Volume 270, 2023, 119947, ISSN 1053-8119, doi.org/10.1016/j.neuroimage.2023.119947; [4]Peng, Han et al., Accurate brain age prediction with lightweight deep neural networks, Medical Image Analysis, Volume 68, 2021, 101871, ISSN 1361-8415, doi.org/10.1016/j.media.2020.101871.

Acknowledgments: This study was partly supported by Helmholtz-AI project BrainAge4AD and funding from TVG.