Do pre-processing and augmentation help explainability? A multi-seed analysis for brain age estimation

Daehyun Cho, Christian Wallraven

Department of Artificial Intelligence, Korea University, Seoul, Korea
1phantasmas@korea.ac.kr
Department of Artificial Intelligence & Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
christian.wallraven@korea.ac.kr

Abstract. The performance of predicting biological markers from brain scans has rapidly increased over the past years due to the availability of open datasets and efficient deep learning algorithms. There are two concerns with these algorithms, however: they are black-box models, and they can suffer from over-fitting to the training data due to their high capacity. Explainability for visualizing relevant structures aims to address the first issue, whereas data augmentation and pre-processing are used to avoid overfitting and increase generalization performance. In this context, critical open issues are: (i) how robust explainability is across training setups, (ii) how a higher model performance relates to explainability, and (iii) what effects pre-processing and augmentation have on performance and explainability. Here, we use a dataset of 1,452 scans to investigate the effects of augmentation and pre-processing via brain registration on explainability for the task of brain age estimation. Our multi-seed analysis shows that although both augmentation and registration significantly boost loss performance, highlighted brain structures change substantially across training conditions. Our study highlights the need for a careful consideration of training setups in interpreting deep learning outputs in brain analysis.

Keywords: Brain Age Estimation · Deep Learning · Explainability · Interpretability · Guided Backpropagation.

1 Introduction

Estimating the age of the brain is essential for detecting abnormalities in brain development, such as neurodegenerative disease or cognitive impairment [20], and has been extensively studied over the past years. As with many other data processing tasks, the advent of deep learning coupled with large, open datasets has significantly increased performance in this domain. These high-performance models, however, suffer from potential overfitting issues given their high capacity and also need to be applied in an explainability framework to open the "black box" [19]. A crucial step for avoiding overfitting has been to use various augmentation strategies that are supposed to increase the robustness and generalizability of the models [30]. Similarly, pre-processing of brain scans, such as anatomical registration, can be done to "help" the models perform better. Explainability methods are then used to create activation maps of those pixels/voxels that are relevant for the model predictions - typically via gradient methods.

We can therefore evaluate deep learning models by their intrinsic metric (i.e., the value of its loss function), but also by their explainability maps (i.e., to what degree do the highlighted regions correspond to known factors of a biological change). In this context it is important to note that training of deep learning models is inherently stochastic due to random weight initialization, dropout, batch size, randomized data augmentations, and stochastic optimization. Hence, running a different "seed" will typically lead to a different sets of weights - even for the same, final loss value. With this in mind, two important open questions remain for gauging the quality of the resulting explanability activations: (i) to what degree are explainability maps consistent across different seeds? and (ii) what effects do pre-processing or augmentation strategies have on explainability?

Here, we present to our knowledge the first, larger-scale study to investigate the effects of seeds, as well as data augmentation and registration in terms of both performance and explainability for the task of brain age prediction. Overall, our contributions are three-fold: first, making use of the stochastic variability across seeds we show that data augmentation results in statistically better (lower) loss compared to non-augmented training for both registered and non-registered brain scans. Second, we investigate the explainability maps via Guided-Backpropagation [31], and find that augmentation results in better-interpretable models, as different seeds share more common voxels. Third, and most importantly, our study uncovers significant changes in explainability already for one deep learning framework across seeds and training setups, highlighting the need for vigilance in interpreting deep learning models for brain age estimation.

2 Related Work

Estimation of age from brain scans evolved from "classic" machine learning regressors to current deep neural networks. As raw voxel images were not suitable for the former models to predict age, brain scans were often processed into features first, which were then regressed onto age [2, 4, 23, 33].

The advent of deep learning and the availability of larger brain datasets also changed performance in this task: convolutional neural network architectures showed substantial improvement in age prediction over the past years - both in 2D (analyzing slices of brains, e.g. [13]) or 3D (analyzing the voxels directly, e.g. [3, 5, 12, 18, 25]).

In addition to improvements in model architecture, data augmentation, in which the model is fed randomly-transformed scans during training, has been shown to improve performance and to avoid overfitting [30] (see [3] for brain age estimation discussion).

Next, explainability has been added to these black-box models [19], since the prediction of age alone is not sufficient for many applications, as practitioners would also like to know or cross-check which parts of the brain actually are involved in aging. Among the many explainability frameworks, GradCAM [29] and Guided Backpropagation [31] are the most widely-used ones, also in the context of age estimation [12].

Another critical aspect of deep learning models is their stochastic nature: various elements of the training are inherently random, which means that different initializations may lead to different models. Recently, this has led researchers to analyze several, randomly-initialized models in their tasks. This can be done to improve performance (ensembling [18]), but also to gauge the statistical robustness of the results [23]. Here, we take the latter approach and launch an in-depth investigation of explainability across different training setups with seeds used to better analyze the inherent variability of the resultant models.

3 Methods

This section describes the dataset and training setups for brain age estimation.

Dataset	#N	Age Mean(±std)	Range
IXI [11]	312	50.37 ± (15.932)	20 - 86
Dallas [24]	273	55.789 ± (19.478)	20 - 89
Oasis1 [22]	315	65.698 ± (9.313)	18 - 94
Oasis3 [17]	552	54.048 ± (21.6965)	42 - 97

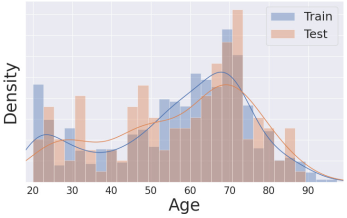


Table 1: Left: Dataset demographics. #N = number of scans. Right: Age distribution of train (orange) and test (blue) sets. Y-axis denotes density.

Dataset: We used a total of 1,452 brain scans from publicly available datasets (see Table 1). To better gauge generalizability, we kept the same 10% (146 scans) as a *hold-out test* set for all runs. During training, the remaining 90% of the brains, were split again in a 90 train/10 validation set ratio (with varying seeds) to introduce variability for each of 100 fixed random seeds.

Preprocessing: All brains were preprocessed starting with skull stripping and normalization through FreeSurfer 6.0 [8]. We then used dipy [7] to register the scans onto the standard MNI152 template including symmetric diffeomorphic registration, removing border voxels outside the brain. All voxels were Min-Max scaled and the scan was cropped to 96 × 96 × 96 voxels.

Augmentation: To investigate the effects of augmentation, we chose a set of the three most popular methods: a random mirroring swapped left and right, whereas random affine (-10 to 10) and random elastic deformations (number of control points 7, max displacement 7.5) were used to further increase anatomical variability. For all augmentations, we used the torchio library [26]. During each mini-batch, one augmentation method was selected at random.

Training setup: We used a standard ResNet50 architecture upscaled to 3D [10], with an Adam optimizer with betas 0.9 and 0.999 and a base-learning rate of 0.0001. The loss was a mean squared error. Given the different data domains, the ResNet was non-pretrained. For each experiment configuration, 100 different runs were done on 2 NVIDIA GeForce RTX 2080 Ti via the same 100 random seeds. Early stopping strategy was applied with 20 epochs of patience triggered by validation data. Multiple model checkpoints were saved to trace performance across training. Overall performance for age prediction was only determined on the hold-out test set. All codes are available at https://github.com/1pha/brain-age-prediction.

Statistical testing: To look for the effects of training setups on performance and epochs, we conducted two-sample t-tests across seeds on the hold-out test set. From each seed, we chose the last and best metrics during the training procedure.

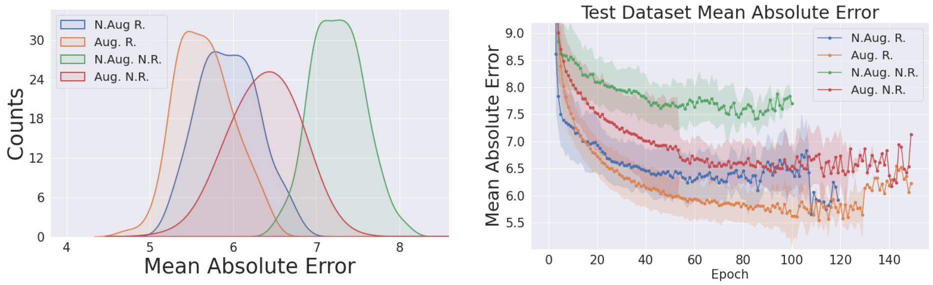
Explainability: In addition to statistical comparisons, we also used the different seeds to analyze the consistency of resultant explainability maps. Given a trained model checkpoint and one brain scan, we applied Guided-Backpropagation (GBP) [31]. The resultant map of gradients from the mean absolute error (MAE) loss with respect to the feature map was then upsampled to the original input brain scan size, $96 \times 96 \times 96$. Obtaining an average explainability map for one checkpoint started with inferring the 146 brains of the test set, followed by GBP, z-normalizing, and then averaging.

As the only "objective" metric to compare models is the loss metric, we gathered checkpoints from seeds once they reached one of five pre-defined loss thresholds (called "Phases", see Table 2 for threshold values), and aggregated their corresponding explainability maps. We retrieved two quantities from the maps: consistent voxels and highlighted regions. To create the former, as the explainability methods assign a higher value to the voxels that influence the prediction, values lower than the 5% quantile values in the aggregated maps were discarded. In order to determine consistently-contributing voxels for predicting age, we chose those that were implicated in more than half of the seeds. The number of agreeing voxels was selected as an important metric for the robustness of the explainability maps. Next, the chosen top 1% percentile values were than aggregated by regions denoted by the AAL ATLAS [27]. Their results were visualized with nilearn [1].

4 Results

4.1 Performance

Figure 1 (a) shows that augmentation overall results in significant improvements in terms of MAE for both non-registered and registered setups across seed distributions. This improvement is also visualized in Figure 1 (b), which shows the evolution of MAE across training for the different setups. As can be expected, however, augmentation also results in significantly longer training of on average 10 to 13 epochs. We note that prediction performance is somewhat lower compared to other works [18, 25], which is likely due to factors including a slightly smaller dataset as well as less aggressive model optimization. Importantly, however, our main objective was to focus on *relative* differences due to augmentation and pre-processing.



(a) Distribution of best test MAE results for all seeds.

(b) Mean and standard deviation of MAE across epochs.

Fig. 1: Mean absolute error statistics on hold-out test dataset. (N.)Aug. = (Non-)Augmented; (N.)R. = (Non-)Registered

4.2 Voxel agreement

We expect that voxels that were repeatedly included in the top-quantiles of the explainability maps across seeds would imply enhanced robustness in explainability. Voxel agreements across conditions are visualized in Figure 2 - see also Table 2 for detailed values. Here, we find that the augmented-registered condition surpassed the other setups. This seems to be a robust finding also throughout training (Figure 2), suggesting that this specific training setup identified reliable voxels early.

For both non-registered conditions (blue curves), the number of agreeing voxels was considerably lower compared to the registered conditions (red curves). This is a result that may be expected since models trained with non-aligned brains would find it harder to localize significant regions.

Similarly, when looking at the effects of augmentation, we also found more agreement in general across seeds (compare solid versus dashed lines), indicating that augmentation aids identification of agreeing voxels. Nonetheless, the "effect size" of augmentation on agreement is lower than that of registration.

4.3 Atlas-based analyses

We next conducted an an atlas-based analysis using the top-1% saliency values from each seeds across training procedures and configurations. In this analysis, we chose to aggregate all top-1% values, regarding the variability across seeds as noise to be averaged out. We then averaged the activated voxels in each of the

Phase	Non-augm./Reg.			Augm./Reg.		
(Threshold)	#C	#E	#L	#C	#E	#L
1 (32)	100	1.76	29.75	100	1.74	29.82
2 (22)	100	2.03	22.55	100	2.02	22.84
3 (7.27)	100	6.13	7.24	100	10.54	7.23
4 (6.0)	53	36.52	5.92	83	36.86	5.97
5 (5.4)	7	53.42	5.34	26	57.03	5.35
Phase	Non-augm./Non-reg.			Augm./Non-reg.		
(Threshold)	#C	#E	#L	#C	#E	#L
1 (32)	100	1.93	30.35	100	1.89	30.45
2 (22)	100	2.19	24.35	100	2.17	23.95
3 (7.86)	98	16.33	7.83	99	10.78	7.83
4 (6.95)	15	48.66	6.92	96	30.05	6.91
5 (6.1)	0	0	0	23	62.0	6.07

Table 2: Information on explainability maps for different checkpoints across training setups. Columns denote as follows: #C number of checkpoints reaching given MAE threshold; #E average training epochs for chosen checkpoints; #L average loss for chosen checkpoints.

atlas-defined brain regions and ranked these across training setup and phases - see Figure 3.

For this type of analysis, the main differences were due to registration setups: the registered conditions mostly implicated subcortical regions including CSF, 3rd & lateral ventricles, or parahippocampal regions. In contrast, non-registered models mostly focus on the occipital lobe - cuneus, occipital gyrus and calcarine fissure. Both setups, however, showed high rankings for the brainstem.

4.4 Region Validation

To check the alignment between the depicted regions from our models and the brain age literature, here we briefly situate our results in the context of published results both in the brain imaging and the deep learning literature.

One of the consistently-nominated brain region from previous research is the *lateral ventricle* along with the *3rd-ventricle*, most notably due to an increase of width [15, 16, 32]. This is clearly matching our results in Figure 3, where these regions for the augmented-registered model have high-rank with minimal deviations.

Similarly we find reports that cerebellar and brainstem volume shrink with age [21], alongside a reduction in thalamic volumes [6,14]. Again, all three regions were implicated in the augmented-registered condition at much higher ranks compared to other training setups.

Ranked regions were also compared with those other published studies on brain age prediction with deep learning. [18] and [12] focus on the lateral, 3rd

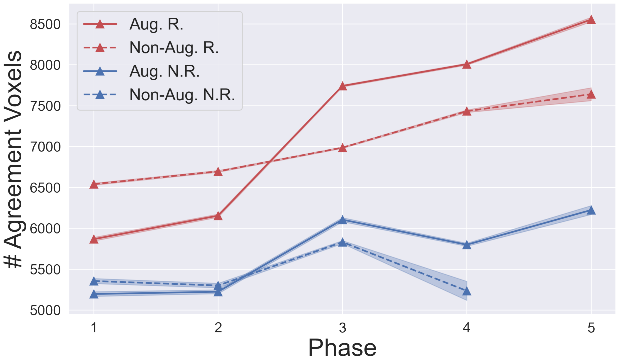


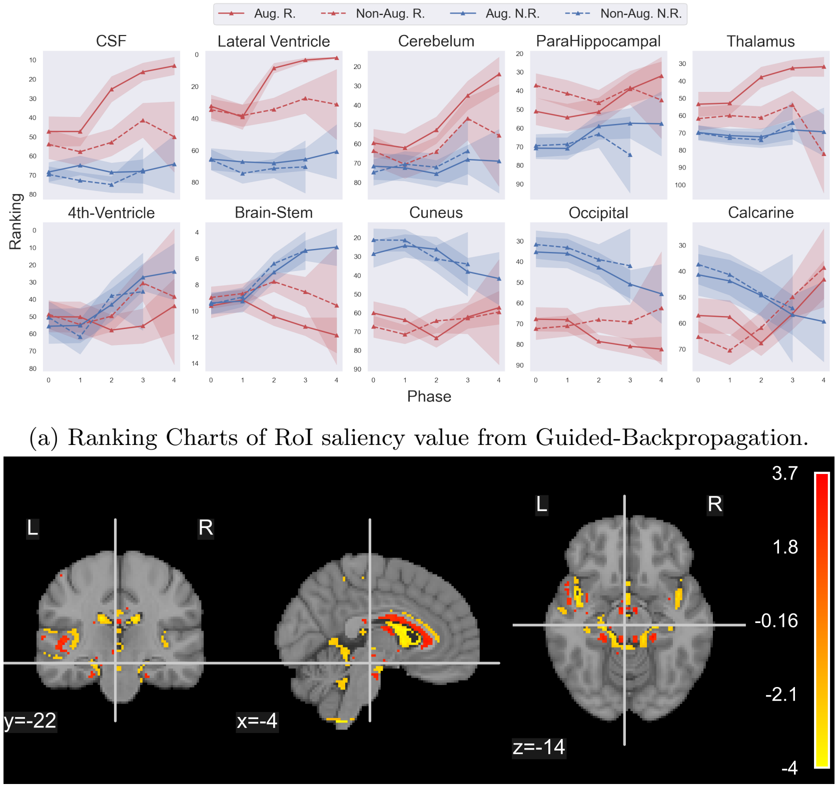
Fig. 2: Evolution of agreement across training. The shaded areas indicate the 95% confidence interval across seeds.

& 4th ventricles and the CSF. The medial temporal structures including the parahippocampals and brain stem are coherent with [28]. [23] mentioned the gray matter cortices, whereas our models mainly focused on the subcorticals and sulci. Overall, significant changes/atrophies in cortical areas could not be found by our work, even with softened thresholds.

5 Conclusion

In our work, we showed that pre-processing - a process that tries to minimize anatomical variability for the same voxel - and augmentation - a process that tries to make the model more sensitive to anatomical variability - both improve the performance of brain age prediction. Importantly, we showed this by means of a larger-scale sample that exposed the intrinsic variability of models with different initialization seeds. Going one step further, we used the resultant explainability maps at well-defined, comparable loss thresholds to trace the evolution of predictive brain regions across training phases and training setups. Here, we found that augmentation on registered brains led to the highest agreement across seeds, suggesting the most robust explainability results for this condition. The brain regions implicated in aging for this training setup focused mostly on previously-implicated subcortical brain areas albeit with little to no activation of cortical areas. Overall, our results highlight the need for caution in interpreting bio-markers for brain age prediction (and, similarly, for other tasks) from only one model - a finding that is also visible in the remarkable variability of brain regions implicated in other studies.

Overall, we believe that our study is an important first step in making more replicable and robust statements about bio-markers from brain scans. Future studies will need to generalize these results with larger brain datasets, test additional deep learning architectures (as for example transformers [9]), as well as compare different explainability frameworks.



(a) Ranking Charts of RoI saliency value from Guided-Backpropagation.

(b) Guided-Backpropagation visualization with augmented-registered condition on the last phase.

Fig. 3: Ranking of region of interest from AAL ATLAS across all seeds.

Acknowledgements This work was supported by the National Research Foundation of Korea under project BK21 FOUR, grants NRF2017M3C7A1041824 as well as by two Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (2019-0-00079, Department of Artificial Intelligence, Korea University; 2021-0-02068-001, Artificial Intelligence Inovation Hub).

References

1. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G.: Machine learning for neuroimaging with scikit-learn. Frontiers in neuroinformatics p. 14 (2014)
2. Aycheh, H.M., Seong, J.K., Shin, J.H., Na, D.L., Kang, B., Seo, S.W., Sohn, K.A.: Biological brain age prediction using cortical thickness data: A large scale cohort study. Frontiers in aging neuroscience **10**, 252 (2018)

3. Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., Montana, G.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage **163**, 115–124 (2017)
4. Dafflon, J., Pinaya, W.H., Turkheimer, F., Cole, J.H., Leech, R., Harris, M.A., Cox, S.R., Whalley, H.C., McIntosh, A.M., Hellyer, P.J.: An automated machine learning approach to predict brain age from cortical anatomical measures. Human brain mapping **41**(13), 3555–3566 (2020)
5. Dinsdale, N.K., Bluemke, E., Smith, S.M., Arya, Z., Vidaurre, D., Jenkinson, M., Namburete, A.I.: Learning patterns of the ageing brain in mri using deep convolutional networks. NeuroImage **224**, 117401 (2021)
6. Fama, R., Sullivan, E.V.: Thalamic structures and associated cognitive functions: Relations with age and aging. Neuroscience & Biobehavioral Reviews **54**, 29–37 (2015)
7. Garyfallidis, E., Brett, M., Amirbekian, B., Rokem, A., Van Der Walt, S., Descoteaux, M., Nimmo-Smith, I.: Dipy, a library for the analysis of diffusion mri data. Frontiers in neuroinformatics **8**, 8 (2014)
8. Greve, D.N., Fischl, B.: Accurate and robust brain image alignment using boundary-based registration. NeuroImage **48**(1), 63–72 (2009)
9. Gupta, U., Lam, P.K., Ver Steeg, G., Thompson, P.M.: Improved brain age estimation with slice-based set networks. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 840–844. IEEE (2021)
10. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6546–6555 (2018)
11. Heckemann, R.A., Hartkens, T., Leung, K.K., Zheng, Y., Hill, D.L., Hajnal, J.V., Rueckert, D.: Information extraction from medical images: developing an e-science application based on the globus toolkit. Proc. 2nd UK E-Sci. Hands Meet (2003)
12. Hepp, T., Blum, D., Armanious, K., Schölkopf, B., Stern, D., Yang, B., Gatidis, S.: Uncertainty estimation and explainability in deep learning-based age estimation of the human brain: Results from the german national cohort mri study. Computerized Medical Imaging and Graphics **92**, 101967 (2021)
13. Huang, T.W., Chen, H.T., Fujimoto, R., Ito, K., Wu, K., Sato, K., Taki, Y., Fukuda, H., Aoki, T.: Age estimation from brain mri images using deep learning. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). pp. 849–852. IEEE (2017)
14. Hughes, E.J., Bond, J., Svrckova, P., Makropoulos, A., Ball, G., Sharp, D.J., Edwards, A.D., Hajnal, J.V., Counsell, S.J.: Regional changes in thalamic shape and volume with increasing age. NeuroImage **63**(3), 1134–1142 (2012)
15. Kaye, J.A., DeCarli, C., Luxenberg, J.S., Rapoport, S.I.: The significance of age-related enlargement of the cerebral ventricles in healthy men and women measured by quantitative computed x-ray tomography. Journal of the American Geriatrics Society **40**(3), 225–231 (1992)
16. Kwon, Y.H., Jang, S.H., Yeo, S.S.: Age-related changes of lateral ventricular width and periventricular white matter in the human brain: a diffusion tensor imaging study. Neural regeneration research **9**(9), 986 (2014)
17. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., Raichle, M.E., Cruchaga, C., Marcus, D.: Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. medRxiv (2019)

18. Levakov, G., Rosenthal, G., Shelef, I., Raviv, T.R., Avidan, G.: From a deep learning model back to the brain—identifying regional predictors and their relation to aging. Human brain mapping **41**(12), 3235–3252 (2020)
19. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
20. Lockhart, S.N., DeCarli, C.: Structural imaging measures of brain aging. Neuropsychology review **24**(3), 271–289 (2014)
21. Luft, A.R., Skalej, M., Schulz, J.B., Welte, D., Kolb, R., Bürk, K., Klockgether, T., Voigt, K.: Patterns of Age-related Shrinkage in Cerebellum and Brainstem Observed In Vivo Using Three-dimensional MRI Volumetry. Cerebral Cortex **9**(7), 712–721 (1999)
22. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience **19**(9), 1498–1507 (09 2007)
23. Niu, X., Zhang, F., Kounios, J., Liang, H.: Improved prediction of brain age using multimodal neuroimaging data. Human brain mapping **41**(6), 1626–1643 (2020)
24. Park, J., Carp, J., Kennedy, K.M., Rodrigue, K.M., Bischof, G.N., Huang, C.M., Rieck, J.R., Polk, T.A., Park, D.C.: Neural broadening or neural attenuation? investigating age-related dedifferentiation in the face network in a large lifespan sample. Journal of Neuroscience **32**(6), 2154–2158 (2012)
25. Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M.: Accurate brain age prediction with lightweight deep neural networks. Medical image analysis **68**, 101871 (2021)
26. Pérez-García, F., Sparks, R., Ourselin, S.: Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Computer Methods and Programs in Biomedicine p. 106236 (2021)
27. Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J., Joliot, M.: Automated anatomical labelling atlas 3. NeuroImage **206**, 116189 (2020)
28. Rueckert, D.: Voxel-level importance maps for interpretable brain age estimation. In: Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data: 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings. vol. 12929, p. 65. Springer Nature (2021)
29. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Gradcam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
30. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**, 60 (2019)
31. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015)
32. Todd, K.L., Brighton, T., Norton, E.S., Schick, S., Elkins, W., Pletnikova, O., Fortinsky, R.H., Troncoso, J.C., Molfese, P.J., Resnick, S.M., et al.: Ventricular and periventricular anomalies in the aging and cognitively impaired brain. Frontiers in aging neuroscience **9**, 445 (2018)

33. Wang, B., Pham, T.D.: Mri-based age prediction using hidden markov models. Journal of neuroscience methods **199**(1), 140–145 (2011)