## Supplementary Information for

# Towards the Interpretability of Deep Learning Models for Human Neuroimaging.

*Simon M. Hofmann*, Frauke Beyer, Sebastian Lapuschkin, Markus Loeffler, Klaus-Robert Müller, Arno Villringer, Wojciech Samek, A. Veronica Witte**

** corresponding authors*

**This file includes:**

Supplementary Methods
Figures S1 to S3
Tables S1
Supplementary References

## Supplementary Methods

### MRI processing stages

***Raw stage*** raw MRI DICOMs were saved in the NifTi-1 format in their native space.

***Freesurfer volume stage*** First, T1-MRIs were subject to the recon-all preprocessing pipeline of FreeSurfer 5.3.0 (Fischl, 2012). After a brain extraction step, intensity normalization procedures, and linear registration to the FreeSurfer standard space, the intermediate processing stage of the T1-image ('brain.finalsurf.mgz') was used to linear register (Rigid, linear interpolation; ANTs 2.2; Avants et al., 2011) also images of the other two sequences (FLAIR, SWI) to the space of the T1-weighted images.

***MNI stage*** To bring images also to a common space across participants, all available sequences were non-linearly warped to the MNI152 space (Fonov et al., 2011) with 2 mm isotropic resolution (ANTs 2.2).

***Freesurfer surface stage*** This stage is a targeted output of FreeSurfer's *recon-all* pipeline, mapping the brain in volume-space to surface-space by creating a 3D-mesh around its folds. The corresponding computed mapping files were later used to first convert and then explore the relevance maps of our interpretation algorithm (LRP) in the individual surface space. Hence, this image stage was only used for visualization and analysis after the training of the prediction models.

### Brain atlases

For the model training on distinct regions of the brain (multi-layer ensemble, MLens type ii), as well as for the structural mapping of LRP relevance distributions, we used a combination of three atlases that cover nearly the entire brain as defined by the MNI152 template: the *Harvard-Oxford* i) cortical and ii) subcortical structural atlases, and iii) the cerebellar atlas (Diedrichsen et al., 2009), all distributed via *FSL 5.0.8*. While there is a minimal overlap between the cerebellum to the other two atlases, we removed the left and right cerebral cortex labels from the subcortical atlas, due to their informational redundancy with respect to the cortical atlas. For details on the Juelich histological atlas, see: *https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases/Juelich.*

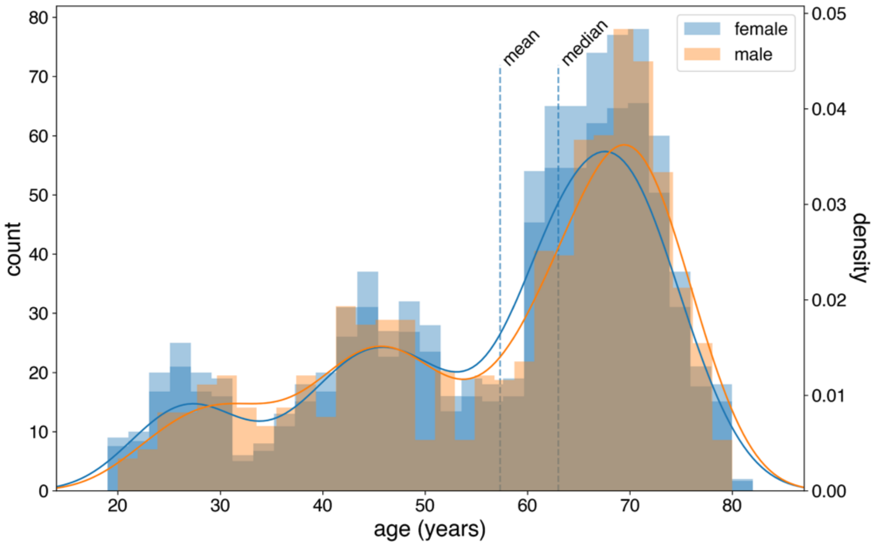


**Fig. S1.** Age distribution in LIFE MRI dataset after exclusion (n=2016, mean age = 57.32, median age = 63.0).

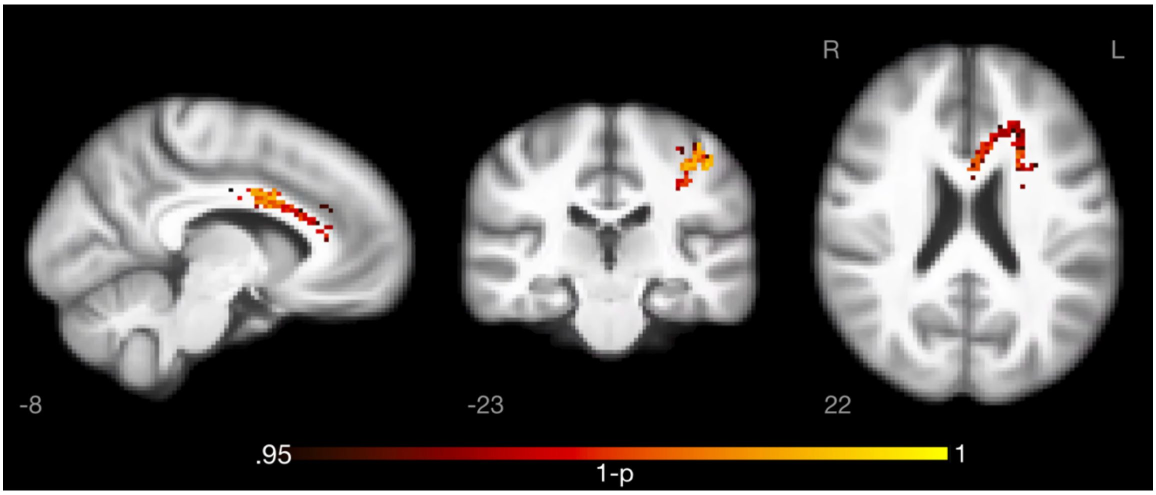


**Fig. S2.** Contrastive relevance maps of diabetics vs. control. For T1-sub-ensemble in MLens type I, subjects with type 2 diabetes mellitus (n=29) were contrasted to controls (n=217; TFCE, FWE-corrected p ≤ 0.05) in older subjects (50-75 years).

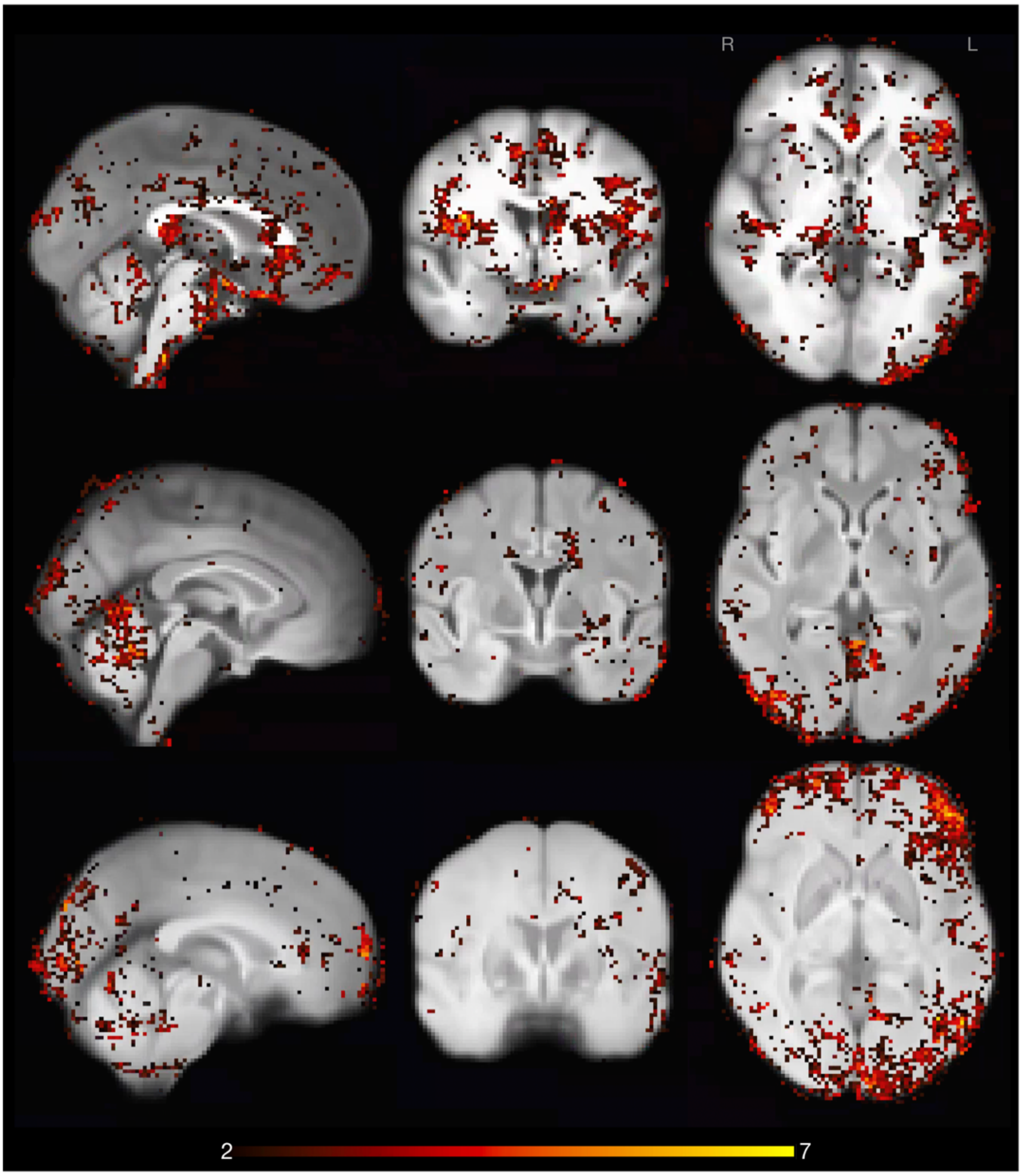
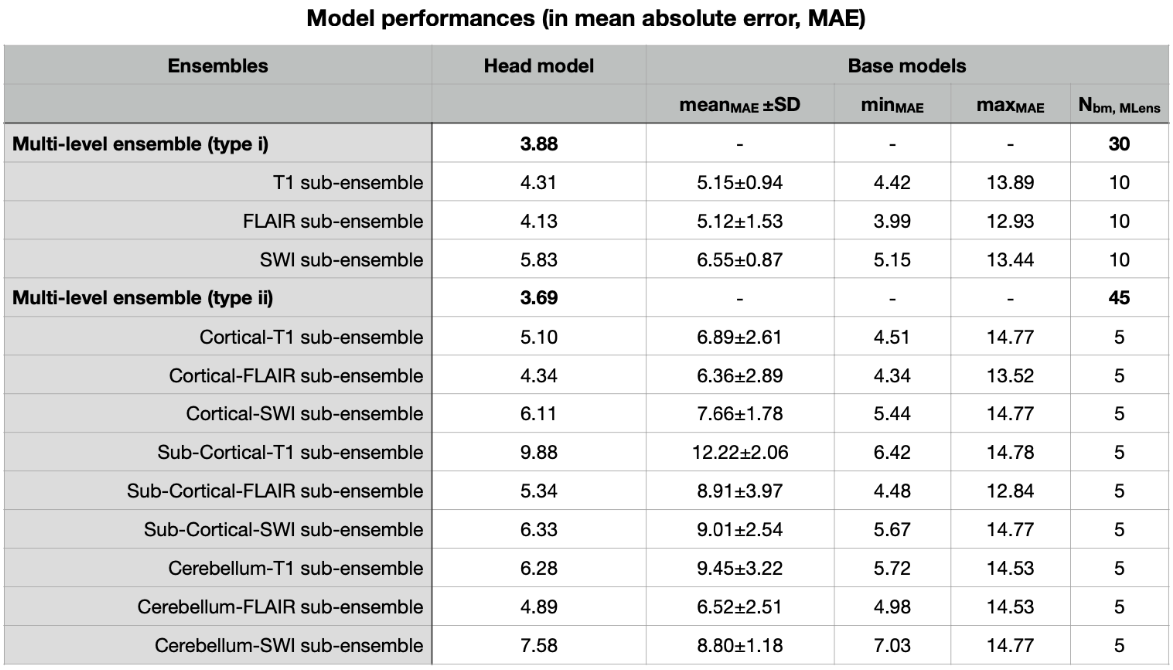


**Fig. S3.** The role of diverging brain age (DBA) on relevance attribution. T-maps (2, 7) of the GLM analysis on the modulation of relevance maps as function of DBA, corrected for age in the older sub-cohort (age ≥ 50)



**Model performances (in mean absolute error, MAE)**

| Ensembles | Head model | Base models | | | |
|---|---|---|---|---|---|
| | | $\text{mean}_{\text{MAE}}$ ±SD | $\text{min}_{\text{MAE}}$ | $\text{max}_{\text{MAE}}$ | $N_{\text{bm, MLens}}$ |
| **Multi-level ensemble (type i)** | **3.88** | - | - | - | **30** |
| T1 sub-ensemble | 4.31 | 5.15±0.94 | 4.42 | 13.89 | 10 |
| FLAIR sub-ensemble | 4.13 | 5.12±1.53 | 3.99 | 12.93 | 10 |
| SWI sub-ensemble | 5.83 | 6.55±0.87 | 5.15 | 13.44 | 10 |
| **Multi-level ensemble (type ii)** | **3.69** | - | - | - | **45** |
| Cortical-T1 sub-ensemble | 5.10 | 6.89±2.61 | 4.51 | 14.77 | 5 |
| Cortical-FLAIR sub-ensemble | 4.34 | 6.36±2.89 | 4.34 | 13.52 | 5 |
| Cortical-SWI sub-ensemble | 6.11 | 7.66±1.78 | 5.44 | 14.77 | 5 |
| Sub-Cortical-T1 sub-ensemble | 9.88 | 12.22±2.06 | 6.42 | 14.78 | 5 |
| Sub-Cortical-FLAIR sub-ensemble | 5.34 | 8.91±3.97 | 4.48 | 12.84 | 5 |
| Sub-Cortical-SWI sub-ensemble | 6.33 | 9.01±2.54 | 5.67 | 14.77 | 5 |
| Cerebellum-T1 sub-ensemble | 6.28 | 9.45±3.22 | 5.72 | 14.53 | 5 |
| Cerebellum-FLAIR sub-ensemble | 4.89 | 6.52±2.51 | 4.98 | 14.53 | 5 |
| Cerebellum-SWI sub-ensemble | 7.58 | 8.80±1.18 | 7.03 | 14.77 | 5 |

**Table S1.** Prediction performances of both multi-level ensembles (type i, ii) with ReLU activation functions and their respective sub-ensembles and 3D-CNN base models (bm), measured in mean absolute error (MAE). Note, MAEs > 13 indicate that a specific model did not learn from the data, i.e., it only output the approximate population mean, which was set at the bias in the output layer.

## Supplementary References

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, *54*(3), 2033–2044. https://doi.org/10.1016/j.neuroimage.2010.09.025

Diedrichsen, J., Balsters, J. H., Flavell, J., Cussans, E., & Ramnani, N. (2009). A probabilistic MR atlas of the human cerebellum. *NeuroImage*, *46*(1), 39–46. https://doi.org/10.1016/j.neuroimage.2009.01.045

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, *54*(1), 313–327. https://doi.org/10.1016/j.neuroimage.2010.07.033