

A voxel-level approach to brain age prediction: A method to assess regional brain aging

Neha GIANCHANDANI

neha.gianchandani@ucalgary.ca

Department of Biomedical Engineering, University of Calgary, Canada

Hotchkiss Brain Institute, University of Calgary, Canada <https://orcid.org/0000-0003-0822-4554>

Mahsa Dibaji

seyedemahsa.dibaji@ucalgary.ca

Department of Electrical and Software Engineering, University of Calgary, Canada

<https://orcid.org/0009-0004-3166-7737>

Johanna Ospel

johanna.ospel@ucalgary.ca

Department of Radiology; Clinical Neurosciences, University of Calgary, Canada

<https://orcid.org/0000-0003-0029-6764>

Fernando Vega

fernando.vega1@ucalgary.ca

Department of Biomedical Engineering, University of Calgary, Canada

<https://orcid.org/0000-0003-0013-8133>

Mariana Bento

mariana.pinheirobent@ucalgary.ca

Department of Biomedical Engineering; Electrical and Software Engineering, University of Calgary, Canada

Hotchkiss Brain Institute, University of Calgary, Canada <https://orcid.org/0000-0001-5125-0294>

M. Ethan MacDonald

ethan.macdonald@ucalgary.ca

Department of Biomedical Engineering; Electrical and Software Engineering; Radiology, University of Calgary, Canada

Hotchkiss Brain Institute, University of Calgary, Canada <https://orcid.org/0000-0001-5421-3536>

Roberto Souza

roberto.souza2@ucalgary.ca

Department of Electrical and Software Engineering, University of Calgary, Canada

Hotchkiss Brain Institute, University of Calgary, Canada <https://orcid.org/0000-0001-7824-5217>

Abstract

Brain aging is a regional phenomenon, a facet that remains relatively under-explored within the realm of brain age prediction research using machine learning methods. Voxel-level predictions can provide localized brain age estimates that can provide granular insights into the regional aging processes. This is essential to understand the differences in aging trajectories in healthy versus diseased subjects. In this work, a deep learning-based multitask model is proposed for voxel-level brain age prediction from T1-weighted magnetic resonance images. The proposed model outperforms the models existing in the literature and yields valuable clinical insights when applied to both healthy and diseased populations. Regional analysis is performed on the voxel-level brain age predictions to understand aging trajectories of known anatomical regions in the brain and show that there exist disparities in regional aging trajectories of healthy subjects compared to ones with underlying neurological disorders such as Dementia and more specifically, Alzheimer's disease. Our code is available at <https://github.com/nehagianchandani/Voxel-level-brain-age-prediction>.

Keywords: Voxel-level brain age prediction, T1-weighted MRI, regional brain aging, deep learning

1. Introduction

As humans progress through life and age, the brain ages as well and it can be observed with neuroimaging (MacDonald and Pike, 2021). This concept, known as brain age, mirrors the chronological age but pertains specifically to the brain. It provides insights into the maturity level and developmental trajectory of an individual's brain which can sometimes be different

from the overall aging process of an individual. For brain age studies, it is assumed that for healthy subjects, brain age is representative of chronological age, indicating that the brain is aging at a similar rate as humans age. However, for subjects with underlying neurological disorders, there is often a deviation in the aging trajectory. An effective biomarker of neurological disorders is increased brain age (Cole et al., 2017, 2018; Huang et al., 2017).

Early works on brain age provide a global estimate, *i.e.*, brain age is studied as a single global index for the entire brain. Global brain age has been demonstrated as an effective biomarker to study the brain aging process in the presence and absence of various neurological disorders (Cole, 2017; Franke and Gaser, 2019). However, due to its global nature, it does not provide spatial information on the brain aging process. Studies have shown that the aging process occurs at different rates and may be non-linear across different regions of the brain, highlighting region-specific response to the aging process (Hof et al., 1996; Raz et al., 2010). The global brain age index is not able to capture this regional information related to aging. The concept of voxel-level brain age can help bridge the gap, where a voxel represents a small unit of the brain volume. Brain age prediction at the level of each voxel can provide a fine-grained analysis of how different regions of the brain age in healthy compared to diseased brains assigning a distinct brain age to each voxel of the brain. Voxel-level predictions can be particularly useful for understanding how neurological disorders impact different regions of the brain. Most neurological disorders are often associated with specific regions of the brain, for example, Alzheimer’s disease (AD) is associated with atrophy in the hippocampus and temporal regions of the brain (Rao et al., 2022; Pasquini et al., 2019), and Parkinson’s is associated with basal ganglia (Blandini et al., 2000; Caligiore et al., 2016), and hence, these regions are expected to have an increased brain age as compared to other regions of the brain in the presence of corresponding disorders.

In this article, an extended analysis and evaluation of our recently proposed deep learning (DL) model to predict voxel-level brain age using T1-weighted magnetic resonance (MR) images (Gianchandani et al., 2023) is presented. The initial work introduced a multitask architecture for voxel-level brain age prediction and evaluation of that model on presumed healthy subjects. In this work, the analysis is extended by performing an ablation study to reflect on how the multi-task architecture is an improvement over a single-task deep learning model. Additionally, the results of the proposed model are inspected and evaluated on subjects with dementia and more specifically, AD and report varying brain ages for different anatomical regions of the brain. A voxel-level brain age prediction model can provide an enhanced understanding of the regional aging processes in the brain while allowing the quantification of the deviation observed in years. Incorporating a multi-task framework moves closer to enhancing the transparency and interpretability of the DL model and it is substantiated by a comparison of the proposed methodology to existing interpretability methods implemented over a state-of-the-art global age prediction model. To summarize, the contributions of this paper are (refer to Figure 1):

1. Proposal of a multitask DL voxel-level brain age prediction model, building upon our prior work (Gianchandani et al., 2023), with an extended evaluation encompassing subjects with dementia and Alzheimer’s disease.
2. An ablation study to show the importance of the different tasks in the multitask architecture.
3. Regional analysis of the brain aging process in presumed healthy subjects and subjects

with dementia and specifically Alzheimer’s disease.

4. Comparison of the proposed model with existing interpretability methods implemented over a state-of-the-art global age prediction model.

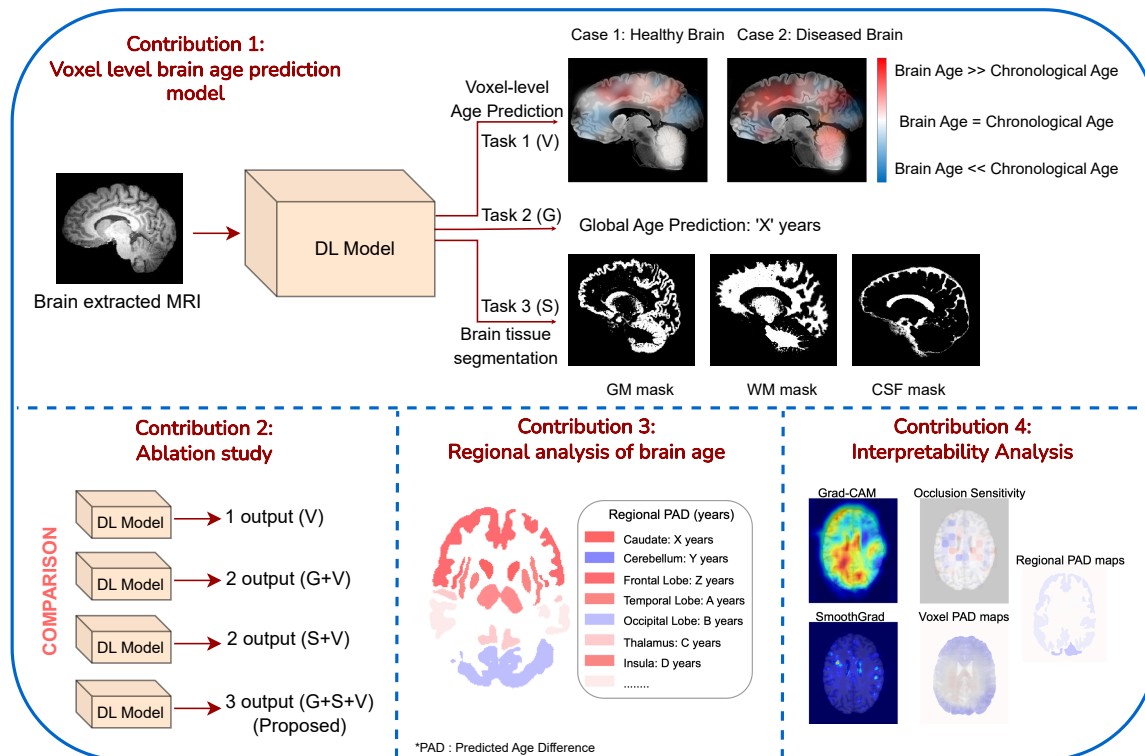


Figure 1: Overview of the contributions of this article. There are 4 major contributions: (i) Proposal of a voxel-level brain age prediction model with initial validation on healthy subjects and subjects with dementia and Alzheimer’s disease (ii) Ablation experiments were done to justify the use of a multitask architecture of the DL model with three-tasks over its two-task, and one-task counterparts. (iii) Regional analysis of predicted brain age by clustering voxel-level brain age predictions into known anatomical regions of the brain and (iv) An interpretability analysis where the proposed voxel-level approach to understanding regional aging trajectories is compared to traditional interpretability methods like Grad-CAM (Selvaraju et al., 2017), SmoothGrad (Smilkov et al., 2017), and Occlusion Sensitivity (Zeiler and Fergus, 2014).

2. Related Work

Brain age prediction is a well-researched domain, however, most studies focus on a global analysis of brain age. Initially, this was done with handcrafted features using traditional machine learning (ML) techniques like Support Vector Machines, Random Forest, and

other traditional machine learning models (Valizadeh et al., 2017; Lemaitre et al., 2012; Beheshti et al., 2021). The approach with traditional ML models is generally considered easier to explain and interpret owing to the reliance on simpler algorithms, fewer parameters, engineered features, and in-built feature importance scores, and achieved brain age predictions with mean absolute error (MAE) \sim 4-8 years. The use of manually-engineered features can aid in understanding the model, but can also be restrictive at the same time as it can lead to the omission of crucial features during the feature engineering process. This limitation led the shift towards the use of DL models for predicting brain age. Manual feature engineering can inadvertently simplify and distort complex data representations, leaving scope for future improvements. Therefore, the transition to neural network models allowed to capture complex data representations within the data that are integral to this brain age prediction task (Plis et al., 2014). DL models showed significant improvement in the brain age prediction task with MAE as low as 2-4 years (Ito et al., 2018; Kolbeinsson et al., 2020), however, due to the neural networks complexity, and black-box nature, these DL models have limited interpretability.

Studies have attempted to explain DL models for brain age prediction with techniques like Grad-CAM (Bermudez et al., 2019), saliency map-based techniques (Yin et al., 2023), occlusion-map based techniques (Bintsi et al., 2021), layer-wise relevance propagation (Hofmann et al., 2022) and SHapley Additive ex-Planations (SHAP) (Ball et al., 2021), among others, to better understand the regional contribution to the brain age prediction models. However, one common limitation of using existing interpretability techniques lies within the use of gradients to calculate feature importance and consequently, the inability to compare the relevance scores across samples. The explanations provided by the existing interpretability methods are quantitative, but only at a sample level as the relevance scores are based on the relative importance of different regions in the input image. Despite the flaws, the aforementioned methods have proven to be tremendously helpful in making the black-box models more transparent and a step closer to understanding the decision-making process of complex neural network architectures. Achieving state-of-the-art results should not come at the cost of interpretability. The proposed approach to predicting voxel-level brain age produces brain predicted age difference (PAD) maps that reflect on the regional aging processes and provide us with a way to quantify healthy versus diseased aging patterns of the brain that is comparable across samples. Additionally, the proposed modeling method ensures that structural features in the brain are used to predict brain age, this will be discussed in detail in Sections 4 and 5.

To move towards a regional analysis of the brain aging process, studies (Beheshti et al., 2019; Bintsi et al., 2020) have attempted to predict brain age at a block or a patch level (with an MAE in the range of \sim 1.5-2 years) where predictions are made for individual blocks of the brain. These blocks do not necessarily correlate to known anatomical regions of the brain but do provide an additional level of spatial information compared to the global-age prediction models. The authors suggest taking this a step further with an analysis at a higher resolution in future works. It is important to acknowledge that studies have attempted to explore and understand regional aging trajectories in the brain using other techniques like regional volume changes (Raz et al., 2005), functional changes (Davidson et al., 1999) etc., however, for the scope of this article, we will be limiting our focus on studies that utilize ML/DL techniques on anatomical images to do so from a brain age prediction perspective.

Finally, based on the current literature, voxel-level predictions have only been explored once before by Popescu et al. (2021). Their method produces voxel-level age maps to understand the regional aging process in the brain, however, this is at the cost of a high MAE ~ 9 years. The authors utilize a modified version of a U-Net architecture to predict brain age at a voxel-level and block-level. This method will be referred to as the baseline for the scope of this article. One common trend observed in most works on brain age prediction (global or regional) is the use of MAE being used as a metric for result comparison. MAE suffers from being influenced by the age range of the test set and hence, makes cross-study comparison less accurate. To overcome this limitation, we do report MAE like previous works, however, also report R^2 (Coefficient of Determination) and show violin plots for the results for a more holistic assessment of results.

Diverging from the baseline (Popescu et al., 2021), the proposed methodology introduces two significant modifications. First, the proposed method models brain age in the native space of the T1-weighted MR images, eliminating registration of any degree as a pre-processing step. This is done to ensure the features are retained in their truest form in the input images. Second, the proposed method uses full T1-weighted volumes as input to the brain age prediction model rather than segmentations of gray matter (GM) and white matter (WM) as done in the baseline. This ensures the inclusion of cerebrospinal fluid (CSF) in the input images which based on previous studies is relevant to the study of brain aging (Houston, 2023; May et al., 1990). The two improvements proposed in this manuscript are further examined in the discussion section.

3. Materials and Methods

3.1 Data

T1-weighted MR imaging is the most widely used MR sequence for brain age prediction (Cole and Franke, 2017; Sajedi and Pardakhti, 2019), likely due to the wide availability of T1-weighted data across a broad age range. Following the same, T1-weighted MR imaging was utilized from publicly available datasets to encourage reproducibility of our work. All data corresponds to presumed healthy controls from the Cambridge Centre for Ageing Neuroscience (Cam-CAN) (Taylor et al., 2017) for training the model. The data was acquired on a Siemens TIM TRIO 3 T scanner. The dataset (n=651) is nearly uniformly distributed across the age range of 18-88 years with a mean age of 54.24 ± 18.56 years. The dataset has a sex-balance of 55%:45%, male:female ratio to limit sex-related bias in the model.

An independent test set (n=359) corresponding to healthy controls for further validation of the model was sourced from the Calgary-Campinas-359 (CC359) dataset (Souza et al., 2018) (age range 36-69 years with a mean of 53.46 ± 9.72 years) with a balanced sex-distribution of 49%:51%. The CC359 dataset contains data acquired on scanners from three different vendors (Philips, General Electric [GE], Siemens) and at two different magnetic field strengths (1.5T, and 3T) giving rise to 6 subsets within the dataset to assess the robustness of the proposed model across different data acquisition protocols.

To create the bias correction methodology (further discussed in Section 3.7), 48 healthy control subjects each from the Open Access Series of Imaging Studies (OASIS) (Marcus et al., 2007), Alzheimer’s Disease Neuroimaging Initiative (ADNI) (Mueller et al., 2005a,b),

and Cam-CAN datasets (unseen during training) were extracted, totalling 144 samples. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The mean age of the bias correction data set was 62.61 ± 21.99 years with a male:female ratio of 50%:50%.

For the evaluation of the proposed model on subjects with underlying neurological disorders, two open-source datasets were utilized. Twenty-eight dementia subjects were extracted from the OASIS dataset (LaMontagne et al., 2019) (mean age 69.17 ± 5.13 years) and twenty subjects with AD from the ADNI dataset (mean age 64.8 ± 5.24 years). The data obtained from the OASIS dataset includes samples with varying dementia types and was chosen to include the analysis of subjects from the broad perspective of cognitive decline. For a more specific analysis, the ADNI dataset was chosen for the analysis of a specific kind of dementia, AD. The OASIS dataset was collected over a period of 10+ years on three different Siemens scanners, (i) Siemens Medical Solutions USA, Inc: Vision 1.5T, (ii) TIM Trio 3T (2 different scanners of this model), and (iii) BioGraph mMR PET-MR 3T. The ADNI samples were extracted from the ADNI 1 cohort which was acquired partly on 1.5T scanners using T1- and dual echo T2-weighted sequences and partly using the same protocol on 3T scanners. Detailed acquisition parameters for the datasets used are described in the source publication for each dataset.

3.2 Data preparation and pre-processing

To ensure that all MR images have the same orientation, FMRIB Software Library’s (FSL) (Jenkinson et al., 2012) ‘fslreorient2std’ command was used. It is important to note that ‘fslreorient2std’ is not a registration command, and hence, no registration is performed, rather it only applies 90, 180, or 270-degree rotations about the different axes as necessary to get the labels of the MR image in the same position as the standard MNI-152 template. Brain extraction masks and tissue segmentation masks to segment GM, WM and CSF for the T1-weighted images from the Cam-CAN dataset were obtained using two U-Net models trained for the specific tasks on the CC359 dataset. The models were trained on the CC359 dataset due to the availability of the binary brain extraction masks and the tissue segmentation masks along with the publicly available T1-weighted images. The binary brain extraction masks are used to obtain brain-extracted input to the model and the tissue segmentation masks are used as ground truths for one of the output tasks in the methodology. All training MR images have a voxel size of 1 mm^3 . Before feeding the MR images to the DL model, random rotation was implemented as an augmentation step. Patches of size $128 \times 128 \times 128$ were cropped from the MR images (1 crop per input image) to be used as input. Random cropping was performed such that the majority of samples have significant brain regions to ensure that the model learns relevant features. The models were implemented using PyTorch with MONAI (Cardoso et al., 2022) for the data preprocessing pipeline.

3.3 Proposed model architecture

In this work, an extended evaluation of our recently proposed multitask U-Net architecture (Gianchandani et al., 2023) is presented to predict voxel-level brain age along with two additional tasks, global brain age prediction and brain tissue segmentation to segment GM,

WM, and CSF. A multitask architecture refers to the presence of multiple outputs that the model is trained for simultaneously. Multi-task learning is known to improve the model training process by including multiple tasks for the model to learn shared representations on, this also helps in avoiding overfitting and leads to fast convergence (Crawshaw, 2020). In the proposed methodology, the main task is the voxel-level brain age prediction task, to complement this task, a brain tissue segmentation task to segment GM, WM, and CSF and a global brain age prediction task are included. Global brain age prediction can be considered a simpler version of the voxel-level brain age prediction task. The segmentation task ensures that relevant structural features are learned from the MR data during training. A good segmentation performance will ensure the learning of structural features like GM, WM, and CSF thickness, shape changes, etc., and owing to the multitask nature of the model, the same features will be repurposed for the brain age prediction task. This ensures the reliance of brain age prediction on structural features in the brain volume. The backbone of the proposed model is a simple U-Net architecture (Ronneberger et al., 2015) that has an encoder and a decoder network, making a U-like shape. Batch-normalization layers are added after the convolution operations to ensure a smooth training process (Santurkar et al., 2018). The encoder and decoder are connected by skip connections that help with recovering important spatial information that is lost during downsampling. The model architecture is depicted in Figure 2.

3.4 Loss function

To accommodate the multitask modeling approach with three different outputs, a custom loss function is defined to ensure that all tasks are given significant importance as the training progresses. The loss function for the proposed model is made up of three terms. $\text{Dice}_{\text{loss}}$ is used to accommodate the segmentation task and is computed from the Dice coefficient (DICE) based on Eq. 1. The Dice coefficient is a measure of the overlap between the ground truth Y and predicted segmentation \hat{Y} . The DICE and $\text{Dice}_{\text{loss}}$ are inversely related, making the model learn accurate segmentations as $\text{Dice}_{\text{loss}}$ is minimized during the training process. $\text{Dice}_{\text{loss}}$ is first computed for all classes *i.e.* background, GM, WM, and CSF individually and then averaged to obtain the overall $\text{Dice}_{\text{loss}}$.

$$\text{Dice}_{\text{loss}} = 1 - \text{Dice} = 1 - \frac{1}{m} \sum_{i=1}^m \frac{2|Y \cap \hat{Y}|}{|Y| + |\hat{Y}|} \quad (1)$$

MAE is the most commonly used metric for the loss function in brain age prediction studies (Feng et al., 2020; Bermudez et al., 2019; He et al., 2021; Popescu et al., 2021). The remaining two terms are two versions of MAE to accommodate the age prediction at the global and voxel-level. Eq. 2 is the voxel-level MAE. First averaged across all brain voxels in the input, followed by batch average, where $y_{i,j}$ is the voxel-level brain age and $\hat{y}_{i,j}$ is the voxel-level predicted brain age for image i and voxel j . Eq. 3 is the global-level MAE, averaged over the batch where y_i is the global brain age and \hat{y}_i is the global predicted brain age for image i . MAE is the absolute difference between the ground truth and the predicted age. In eqs. (1) to (3), m is the batch size, and n is the total number of brain voxels in one sample.

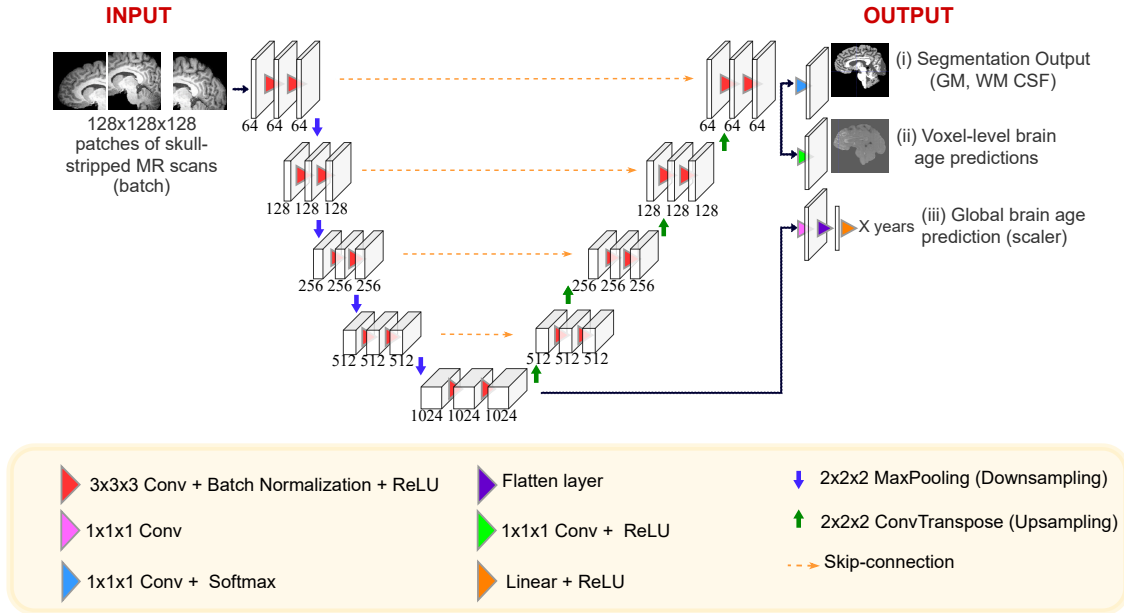


Figure 2: The proposed multitask model used for voxel-level brain age prediction. The model follows a U-Net backbone with a downsampling and upsampling block to obtain the output at the same size as the input. It takes cropped MR patches of size $128 \times 128 \times 128$ as input to the model and produces three outputs, a voxel-level brain age prediction *i.e.* the main output as well as two secondary outputs, brain tissue segmentation of GM, WM and CSF, and global brain age prediction.

$$\text{MAE}_{\text{voxel}} = \frac{1}{m} \sum_{i=1}^m \frac{1}{n} \sum_{j=1}^n |y_{i,j}^{\text{noise}} - \hat{y}_{i,j}| \quad \text{where } y_{i,j}^{\text{noise}} = y_{i,j} + U(-2, 2) \quad (2)$$

$$\text{MAE}_{\text{global}} = \frac{1}{m} \sum_{i=1}^m |y_i - \hat{y}_i| \quad (3)$$

The weighted sum (Eq. 4) of the three terms is the loss function (\mathcal{L}) to be optimized during training. The weights w_s , w_g , and w_v were set empirically and changed as the training progressed. The weight for the segmentation output, w_s , was initialized with the highest weight owing to the value being the smallest among the three loss terms (ranging between 0-1). The global age and voxel-wise age prediction weights, w_g , and w_v , respectively, were initialized with equal weights and updated as described in Table 1. The individual weights were manually initialized and changed as the training progressed rather than being learned during training as hyperparameters to ensure the weights do not simply converge to zero to minimize the overall loss without effectively minimizing the individual loss terms. The choice of weight initializations were inspired by the range of each of the loss metrics as well as

Table 1: Loss function weights as the training progresses.

Weight	Epochs ∈ [0, 50)	Epochs ∈ [50, 130)	Epochs ∈ [130, 300]
w_s	80	40	15
w_g	1	1	0.7
w_v	1	1	1.3

observation of the individual loss values in the initial epochs as the training progressed. This technique included experimentation to arrive at the weights proposed in the manuscript.

$$\mathcal{L} = w_s DICE_{loss} + w_v MAE_{voxel} + w_g MAE_{global} \quad (4)$$

To ensure that the model does not learn a uniform prediction of brain age across all voxels, a subtle noise component is introduced into the loss calculation during the model’s training. This uniform noise (U), randomly selected from the range of -2 to +2, is added to the ground truth labels for each voxel (Eq. 2). This strategic addition of noise encourages the model to learn the nuanced variations in the brain aging process across distinct regions and across subjects. We hypothesize that, when exposed to a combination of added noise and variations in underlying structural features, the model will develop accurate representations of these variations during training. By constraining the noise to a narrow range of -2 to +2, it is ensured that the effect on the training process is limited to an intentionally added randomization, without significantly impacting the model training process. The noise component influences the MAE calculation during the training loss computation and consequently, helps with a robust update of weights during back-propagation. The noise, despite being small, helps the model to learn variations in structural features as observed in the T1-weighted images by acting as a tool to guide the model towards learning features such that all voxels are not assigned the same brain age predictions, but rather, predict brain age based on small variations in structural changes as observed in the T1-weighted images.

To evaluate the significance of incorporating noise into the ground truth labels, a variant of the proposed model without the inclusion of any additional noise was also trained. In this model, chronological age is assigned to each voxel in the ground truth labels for the voxel-level brain age prediction task based on the assumption discussed in Section 1. In the subsequent result section, a performance comparison of both models is described.

3.5 Ablation study

An ablation study was performed to verify the choice of a multitask architecture. The objective is to demonstrate that both the global-brain age prediction task and the brain tissue segmentation task contribute to the model learning enhanced and accurate representations, specifically geared towards improving performance in the primary task *i.e.* the voxel-level brain age prediction. Multiple models are trained, starting with a single output model that predicts voxel-level brain age, iteratively adding the other two tasks, one at a time, to analyze how models with different output tasks trained on the same dataset perform in

comparison to one another. Thus, 4 different models were trained: 1) a one-task model to predict voxel-level brain age (V), 2) a two-task model to predict voxel-level brain age and segmentations of GM, WM and CSF (S+V), 3) a two-task model to predict voxel-level brain age and global-level brain age (G+V) and 4) a three-task model that predicts voxel-level brain age, global-level brain age and segmentations of GM, WM and CSF (proposed model, S+G+V).

3.6 Network training

3.6.1 PROPOSED MODEL

The Cam-CAN dataset was used for training the proposed model. A train:validation:test split of 489:64:98 subjects was used. Patches of size $128 \times 128 \times 128$ voxels were randomly cropped from the MR images on the fly and used as input to the model. Using patches is helpful in reducing the computational load during training allowing for the incorporation of a bigger batch size. Random cropping was done to ensure that a large majority of data samples in each batch had a significant part of the brain region, and randomizing the cropping process helps in exposing the model to brain regions from different perspectives, leading to accurate and robust features being learned. The model was trained for 300 epochs with a batch size of 2. The Adam Optimizer was used with an initial learning rate of 0.001, weight decay of $1e-5$, and beta values set to (0.5, 0.999). The learning rate decreased every 70 epochs by a multiplicative factor of 0.6. The hyperparameters were empirically selected.

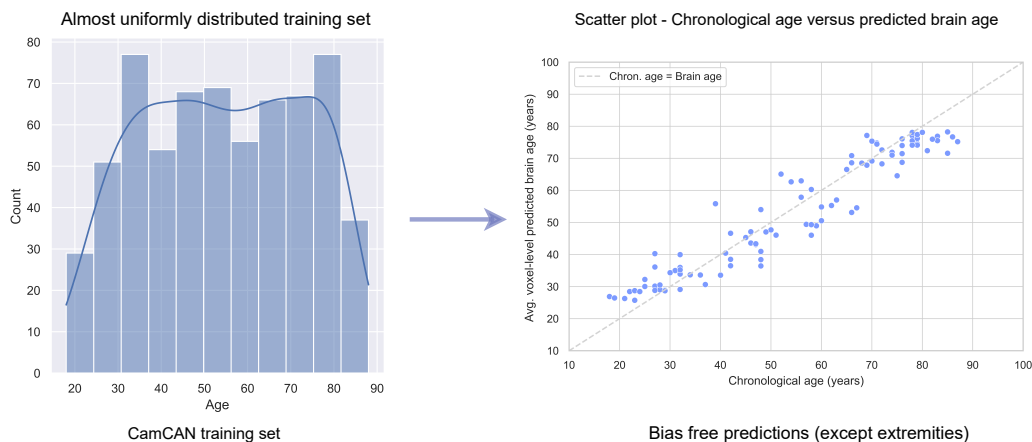


Figure 3: (left) Cam-CAN training set follows a rough uniform data distribution, exposing the proposed model to samples of all ages. (right) This leads to bias-free predictions mostly, except for the extremities (ages 20-30 and ages 80-90). It can be observed that the predictions are closely aligned around the regression line for ages 30-80, with slight bias observed on the edges. A correction methodology can help correct the observed bias.

3.6.2 ABLATION STUDY

The same train:validation:test split of the Cam-CAN dataset used for the proposed model was used to train the ablation study models described in Section 3.5. All ablation experiment models were tested on 50 test set subjects from the Cam-CAN dataset and 359 subjects from the CC359 test set. The CC359 was split into six subsets (as described in Section 3.1) based on the scanner used and the magnetic field strength at which the data was acquired. Metrics were obtained for each of the six subsets to compare performance across the varying subsets.

The one-task model to predict voxel-wise brain age and the two two-task models (segmentation/global age + voxel-wise brain age) were all trained for 300 epochs. Various hyperparameters were experimented with. However, the most suitable ones were found to be similar to the ones used to train the proposed model, with a slight difference in the beta values that were set to default (0.9, 0.999) for the Adam optimizer.

3.7 Bias Correction

Bias correction is a post-processing step in brain age prediction pipelines. This step is essential to remove bias due to the mean age of the training set. Brain age prediction models have been observed to be biased around the mean age of the training dataset, leading to under-estimations of brain age for subjects older than the mean age and over-estimations for subjects younger than the mean age. The source of this bias is largely unknown but is speculated to be due to reasons including noisy data, heterogeneity in the training set, data distribution, availability of data corresponding to varying age ranges, and the modeling techniques used (Aycheh et al., 2018; Cole et al., 2017; Liang et al., 2019). A uniform dataset (Cam-CAN) during training was used, exposing the model to a balanced number of samples across all age ranges (and balanced sex distribution), minimizing biased predictions. However, despite using an ostensibly uniform dataset, the number of samples in the extremities (20-30 years, and 80-90 years) is comparatively lower than the rest.

The proposed methodology employed for the bias correction technique followed what was proposed in Popescu et al. (2021), which based on the current literature is the only study that proposed a bias correction for voxel-level brain age prediction algorithms. Hence, the goal is to train a model that learns age-specific structural features relevant to predict age such that the predictions have minimal bias. This can be confirmed by comparing the results before and after bias correction, a small difference between the two indicates that bias correction does not impact the results significantly, and hence, predictions are minimally biased.

3.8 Regional Analysis of PAD maps

Research in the field of brain aging studies the aging process at a regional level *i.e.* in the context of different regions of the brain. To better understand the PAD maps and to assess the clinical relevance, a regional analysis of the predicted age difference at the level of known regions of the brain was performed. The publicly available MNI structural atlas (Collins et al., 1995; Mazziotta et al., 2001) provided by the Research Imaging Center, University of Texas Health Science Center at San Antonio, Texas, USA that segments the brain into

9 anatomical regions namely Caudate, Cerebellum, Frontal Lobe, Insula, Occipital Lobe, Parietal Lobe, Putamen, Temporal Lobe and Thalamus is used. Additionally, regional averages for the Ventricles and White Matter in the brain are also computed in the MNI space. To obtain the said averages, the MNI152 (Fonov et al., 2011) template was used to obtain the Ventricle and White Matter segmentation using the “SynthSeg” tool (Billot et al., 2023) available as part of the FreeSurfer software package (Fischl, 2012). Voxel-level brain PAD values are aggregated within each of the 11 regions to compute the average brain PAD for each region in the healthy and diseased test sets.

3.9 Overview from an interpretability perspective

Previously, with the aim of understanding regional contributions to brain age and ensuring accurate features are learned during training, global age prediction models have been explained using traditional interpretability methods. In this contribution, insights obtained from the voxel-level PAD maps are compared to the ‘traditional’ way of understanding the models. To do so, a publicly available state-of-the-art Simple Fully Convolutional Neural network (SFCN) for global age prediction (Peng et al., 2021; Gong et al., 2021) was used and three interpretability methods were implemented on it: (i) Grad-CAM (Selvaraju et al., 2017), (ii) Occlusion Sensitivity maps (Zeiler and Fergus, 2014) and (iii) SmoothGrad (Smilkov et al., 2017). The heatmaps/saliency maps obtained were contrasted against voxel-level and regional-level PAD maps and observations were discussed.

The SFCN model was originally designed to approach the brain age prediction task as a soft classification task, however, for the proposed implementation, the output layers of the architecture are modified to a regression head and same feature extractor is utilized as done in the original work. The Cam-CAN dataset was used to train the model following the same train:test split as done for the proposed model for fairness with the difference lying in the preprocessing of the input MR images. As the original modeling process utilized linearly registered images, the same steps were performed to linearly register the training images to the MNI template before feeding them as input to the model. An important consideration here is that no registration is performed for the proposed model, and hence the PAD maps obtained are in the native image space, whereas the interpretability heatmaps obtained are in the MNI space. Even though linear registration (or 6 degrees of freedom registration) does not alter the shape of the brain as it only implements translational and rotational changes, it can still introduce smoothening in the brain features during the process, which we want to avoid to retain the structural integrity of the brain. On the other hand, non-linear registration can distort the existing brain structure in aligning the MR scan to the MNI space, which is also not desired. The uniqueness of each brain’s shape and structure contributes to the prediction of brain age, and hence, it was decided against performing any registration (linear or non-linear) for the voxel-level brain age prediction model.

4. Results

For a fair comparison of model performance and as suggested in Popescu et al. (2021), all results are reported before bias correction. Bias-corrected results are only used for visualizations and analysis of diseased subjects where explicitly stated. In Table 2, columns 2 and 3 with header voxel-level MAE refer to averaged voxel-level prediction results, and

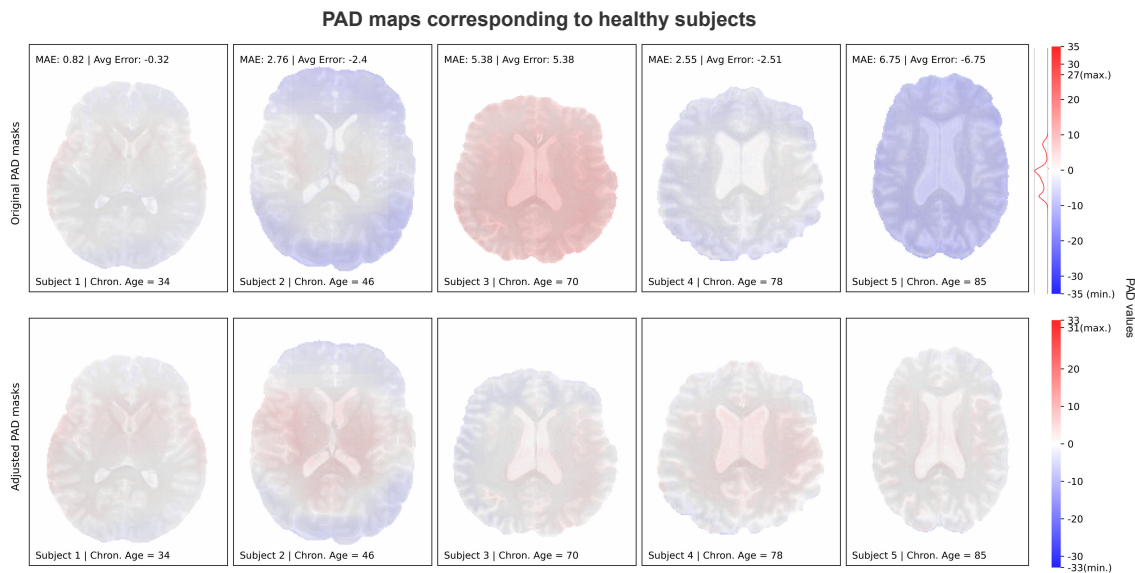


Figure 4: Row 1 - PAD maps (raw, with no bias correction) based on the voxel-level difference between chronological and predicted age, Row 2 - adjusted PAD maps by subtracting the overall MAE of the brain volume from each voxel PAD value. Each row is plotted with an independent colormap based on the range of values observed in the samples plotted in that row. Minimum and maximum points on the color bar denote the minimum and maximum voxel PAD observed across the samples in the specific row. The data distribution plot beside the color bar in row 1 shows the distribution of PAD values across the entire healthy test set. Extended analysis of healthy PAD maps is described in (Gianchandani et al., 2023).

columns 4 and 5 with header global MAE refer to the global brain age prediction results. Voxel-level MAE results are considered for model comparisons throughout the work.

Contribution 1: Proposal of a multitask DL voxel-level brain age prediction model: The proposed model surpasses the baseline (refer Table 2 - Voxel-level MAE), demonstrating a 39.22% reduction in MAE on the internal Cam-CAN test set. The proposed model is also evaluated on a larger external test set (CC359) and obtains an MAE reduction of 58.88% which reflects on the model’s performance on unseen data originating from a different data source. The proposed model variant (with 3-output) without added noise to the loss function comes in second on the Cam-CAN evaluation and second to last on the CC359 test set. The global level MAE is reported for the models that included global brain age predictions in Table 2 - Global MAE column. It can be observed that the SFCN model outperformed the remaining models on the Cam-CAN test set, whereas the proposed model achieved the smallest MAE on the CC359 test set, however, it must be noted that this work aims to predict voxel-level brain age and global brain age was added as a complementary task in the proposed model for improved feature extraction for the main voxel-level brain age prediction task. Violin plots showing the data distribution of voxel-level test results can

be observed in Figure. 5. The median MAE is seen to be lower for the proposed model as compared to the baseline on both test sets. It can also be observed that variation in model performance is almost comparable for all models, however, for both test sets, the widest part of the violin that indicates higher variability or spread in MAE is seen at a lower error value for the proposed model as compared to the baseline. Figure 6 shows the scatter plots for the baseline and proposed model’s voxel-level brain age prediction results with respective R^2 scores (coefficient of determination) that show the goodness of fit of a model.

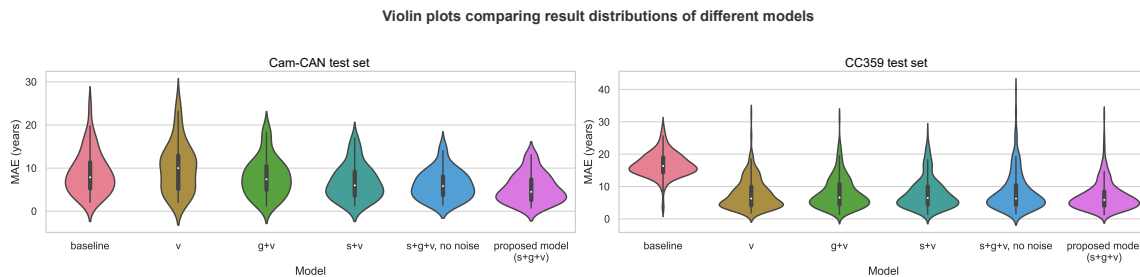


Figure 5: Violin plots comparing the distribution of test results (voxel-level MAE) of different models. Each violin represents the probability density of data, with quartiles displayed as black lines and the median indicated by the white dot. It can be observed that for both the Cam-CAN and CC359 test sets, the median MAE is smaller for the proposed model compared to the baseline. All results are reported before bias correction.

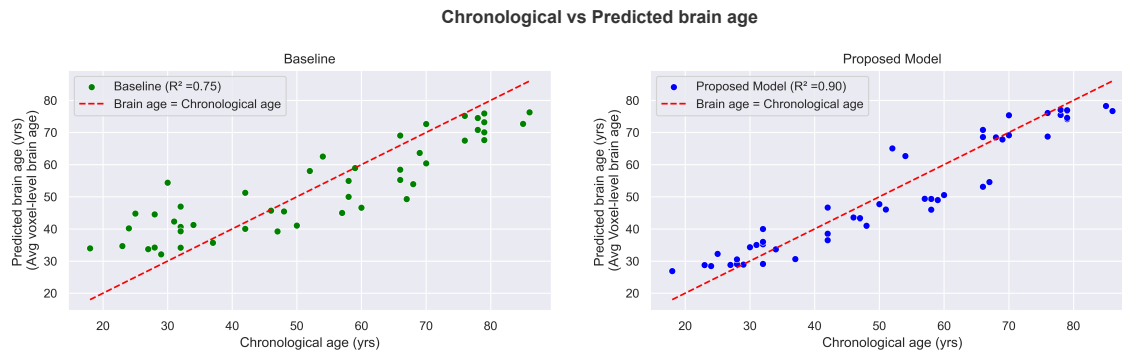


Figure 6: The scatter plots show the predicted versus chronological age for the baseline and proposed model. The baseline achieved an R^2 of 0.75 whereas the proposed voxel-level brain age prediction model achieved an R^2 of 0.90 on the Cam-CAN test set.

For voxel-level predictions, since it is impossible to present prediction results at the level of each voxel (millions in each brain volume), the mean of the per-sample MAE (MAE_{voxel}) is reported in Table 2. To visualize the voxel-level brain age predictions, predicted age

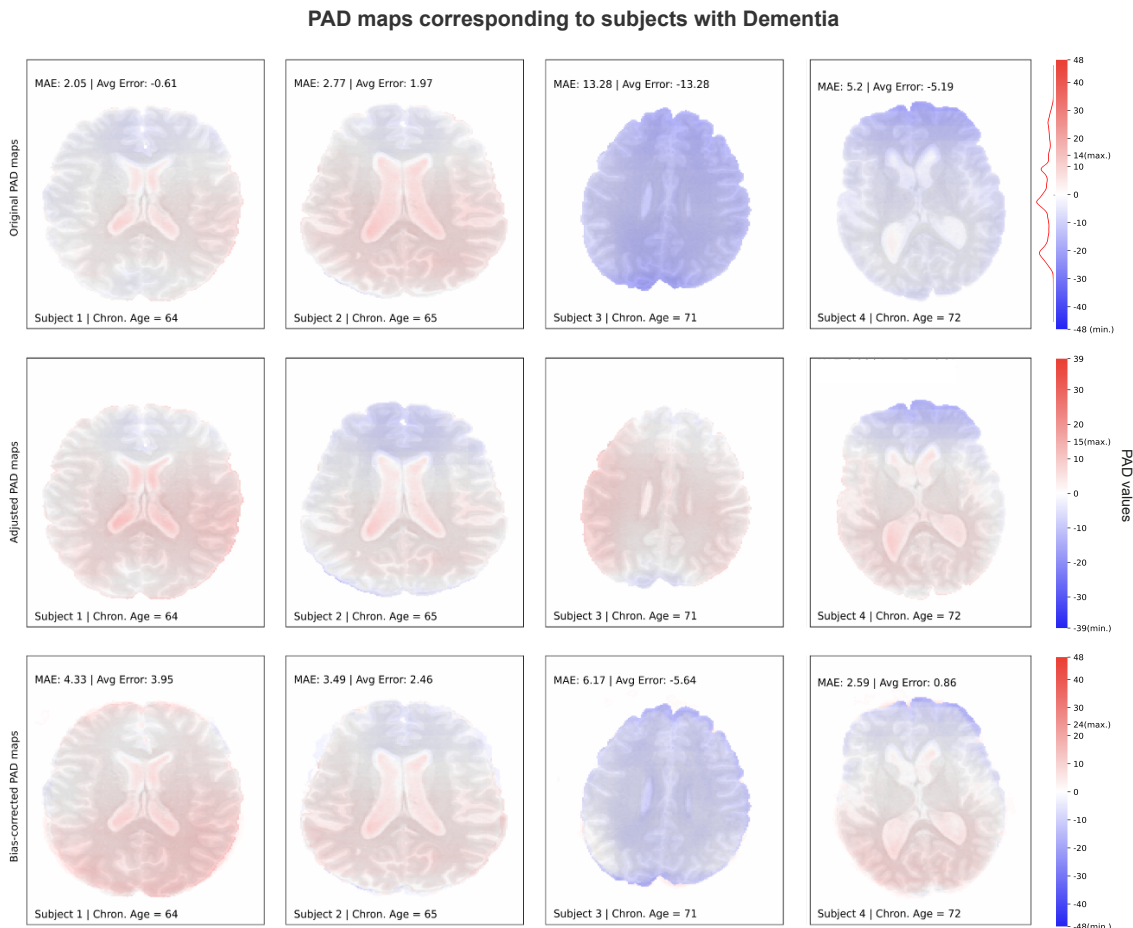


Figure 7: PAD maps corresponding to diseased subjects (OASIS dataset). Row 1 shows the raw PAD maps obtained from the voxel-level brain age prediction model. Row 2 shows the adjusted PAD maps (for improved visualization). Row 3 shows bias-corrected PAD maps using the correction methodology described in Section 3.7. More red regions are observed as compared to healthy PAD maps and accelerated aging in the ventricles which has often been associated with neurological disorders. Each row is plotted with an independent colormap based on the range of values observed in the samples plotted in that row. Minimum and maximum points on the color bar denote the minimum and maximum voxel PAD observed across the samples in the specific row. The data distribution plot beside the color bar in row 1 shows the distribution of PAD values across the entire test set of subjects with dementia. A wider spread of PAD values is observed compared to the data distribution of the healthy test set.

difference (PAD) maps are used, which show the difference between the predicted brain age and the chronological age at the level of each voxel. PAD maps for the Cam-CAN

Table 2: Model performance on an internal (Cam-CAN) and external test set (CC359). All results are reported before bias correction.

Model (output tasks)	Voxel-level MAE		Global MAE	
	Cam-CAN	CC359	Cam-CAN	CC359
Global age (G)	-	-	5.32±3.67	6.50±4.71
Baseline (G+V)	8.84±4.82	16.74±3.71	-	-
1 output model (V)	10.11±5.68	7.63±4.53	-	-
2 output model (G+V)	7.90±4.30	7.93±4.73	6.61±4.17	6.52±5.22
2 output model (S+V)	6.75±3.94	7.83±4.74	-	-
3 output model (S+G+V), no noise	6.14±3.32	8.32±5.84	5.83±3.98	8.70±7.16
Proposed model (S+G+V)	5.30±3.29*	6.92±4.28*	6.11±3.80	5.51±4.38

Abbreviations: V - voxel-level brain age prediction task, S - segmentation task (GM, WM, CSF), G - global-level brain age prediction task, * - $p < 0.05$

test set samples can be observed in Figure 4, where blue color indicates brain regions that look younger than chronological age and red correlates to older-looking brain regions. The first row corresponds to the raw PAD maps whereas the second row corresponds to the adjusted PAD maps obtained by subtracting the overall MAE of the brain volume from each voxel PAD value. These adjusted maps allow us to visualize the spatial variations in PAD values across different regions of the brain without the interference of the model error (MAE). The adjusted PAD maps are constructed purely for visualization purposes and are not used for any result comparisons with other models/baseline. Similarly, the PAD maps corresponding to subjects with dementia can be observed in Figure 7. At a high level, it can be observed from the PAD maps corresponding to healthy versus dementia subjects, that the contrasts are sharper and more apparent in subjects with dementia reflecting greater variation in regional brain ages. Additionally, the PAD maps for subjects with dementia have intensity PADs spread across a wider range of values, which can be observed from the distribution of values shown alongside the color bar in row 1 in Figure 7 as well more red regions as compared to healthy PAD maps. More analysis on healthy PAD maps is done in Gianchandani et al. (2023) and that on diseased subjects will be further discussed in the subsequent sections.

The Wilcoxon-Signed Rank test was performed to assess the voxel-level performance of the proposed model against other variations (1-output, 2-output) of the model and the baseline. Hence, 5 statistical tests were performed, the proposed voxel-level brain age prediction model against each of the models in rows 2-6 of Table 2. α was set to 0.05 and the Holm-Bonferroni correction was done to account for multiple comparisons. All resulting p-values were found to be less than 0.05, indicating statistical significance.

The proposed model was also tested on the CC359 dataset stratified by sex. The model performed with a difference of ~ 1.2 year in the voxel-level MAEs on the two test sets. The proposed model achieved an MAE of 7.54 ± 4.78 years on the CC359_{male} test set (n=176) and an MAE of 6.32 ± 3.61 years on the CC359_{female} test set (n=183).

Table 3: Ablation study results. All results are reported before bias correction.

Test Set	Model (output)	MAE±S.D.
Philips 1.5T	V	7.22±3.13
	S+V	7.83±4.63
	G+V	9.20±5.36
	S+G+V (proposed)	6.94±3.80 * S+V,G+V
Philips 3T	V	8.02±5.29
	S+V	9.61±5.33
	G+V	9.54±5.99
	S+G+V (proposed)	7.73±5.04 * S+V,G+V
Siemens 1.5T	V	8.26±5.33
	S+V	8.64±5.60
	G+V	8.75±4.37
	S+G+V (proposed)	6.68±4.80 * V, S+V,G+V
Siemens 3T	V	9.18±5.29
	S+V	6.21±4.13
	G+V	5.84±4.10 * V, S+G+V
	S+G+V (proposed)	6.80±4.22
GE 1.5T	V	5.83±3.21
	S+V	7.17±3.51
	G+V	5.79±2.34 * S+V
	S+G+V (proposed)	5.98±2.52
GE 3T	V	7.26±3.46 * G+V
	S+V	7.55±4.08
	G+V	8.49±3.73
	S+G+V (proposed)	7.40±4.52

1. All test sets have n=60 samples, except Philips 1.5T with n=59 samples.
2. Abbreviations: V - voxel-level brain age prediction task, S - segmentation task (GM, WM, CSF), G - global-level brain age prediction task.
3. Statistical significance is shown with the symbols, V, S+V, G+V or S+G+V which denote each model. The symbol beside an MAE value indicates statistical significance (with $p<0.05$) with the specific model mentioned.

Contribution 2: Ablation study to show the importance of using a multitask architecture: As stated in Section 3.5, the proposed three-task (multitask) model is expected to show superior performance on the voxel-level brain age prediction task compared to the one-task and two-task counterparts. An ablation study is performed by designing experiments with the same model architecture with different task combinations, and it can be observed in Table 2, that the 3-output proposed model outperforms the 1-output and 2-output models with statistically significant results ($p<0.05$) on the internal Cam-CAN test set. To further validate the findings, all ablation study models are subjected to evaluation

using the CC359 dataset. This dataset comprises data acquired from 3 distinct scanner vendors, each acquired at 2 different magnetic field strengths. Consequently, this dataset is segregated into 6 subsets, all sharing similar acquisition protocols. The evaluation is conducted independently on each subset (refer to Table 3) for every ablation experiment model. It is observed that the proposed model outperforms the 1-output (V) and 2-output models (S+V, G+V) on 3 out of 6 subsets (Philips 1.5T, Philips 3T, Siemens 1.5T), comes close second on 1 subset (GE 3T) and takes the third spot on the 2 subsets (GE 1.5T, and Siemens 3T). Closely inspecting the subsets where the proposed model did not take the lead, it was observed that for the GE 3T subset, the proposed mode ranked second with an average MAE on the test set differing by no more than 0.2 years. Similarly, on GE 1.5T and Siemens 3T subsets, where the proposed model secured the third position, the difference between the top-ranking model and the proposed three-task model was at best 1 year.

The Wilcoxon-Signed Rank test was performed to assess the statistical significance of the results. For each subset of CC359, the winning model was compared to the remaining three models, and multiple comparisons were accounted for by performing the Holm-Bonferroni correction. For each test set, if the winning model was found to have significant results against another model, the symbol assigned to the model has been shown beside the MAE value in column 3. It can be observed that even though the proposed model obtained significantly better results against all the other model variants only on the Siemens 1.5T test set, the performance is consistently better than any other variants of the proposed model. The 1-output (V) and 2-output (S+V, G+V) models only performed significantly better than 1 or 2 other models whereas the proposed model consistently had significantly better results than 2 or 3 of the other model variants.

Overall, the proposed model outperformed the ablation experiment models on 50% of the subsets, while consistently performing well across all subsets, unlike the 1-output and 2-output models which obtained significantly higher errors (~ 9 years) on at least 1 or more of the subsets. The proposed model consistently achieved an average MAE in the range of 5.9 to 7.7 years across all subsets of CC359, whereas other ablation experiment models (1-output and 2-output) exhibited greater fluctuations in the inter-dataset performance. Evaluation on subsets acquired using different scanners, which in turn exhibit scanner-specific differences in the MR images, and at different magnetic field strengths reflects on the model’s ability to be robust and generalizable across diverse datasets.

Contribution 3: Regional analysis of the brain aging process in a healthy versus diseased brain: The proposed model was tested on healthy subjects from the Cam-CAN dataset, which was used for the regional analysis. For the evaluation of diseased subjects, subjects with AD from the ADNI dataset (n=20) and subjects with dementia from the OASIS3 dataset (n=28) were utilized. It is essential to note that the majority of the open-source MR images of subjects with neurological disorders (especially AD and dementia) correspond to older age ranges, usually 55 years and above with the frequency of samples available increasing as one goes higher up. To mitigate any biased predictions, filtering was performed on all test sets (healthy, AD, and, dementia) for subjects with age ≤ 70 years for the regional analysis, leaving us with n=40 healthy subjects and n=32 subjects with either AD or dementia for the analysis. This decision will be further justified in the discussion section.

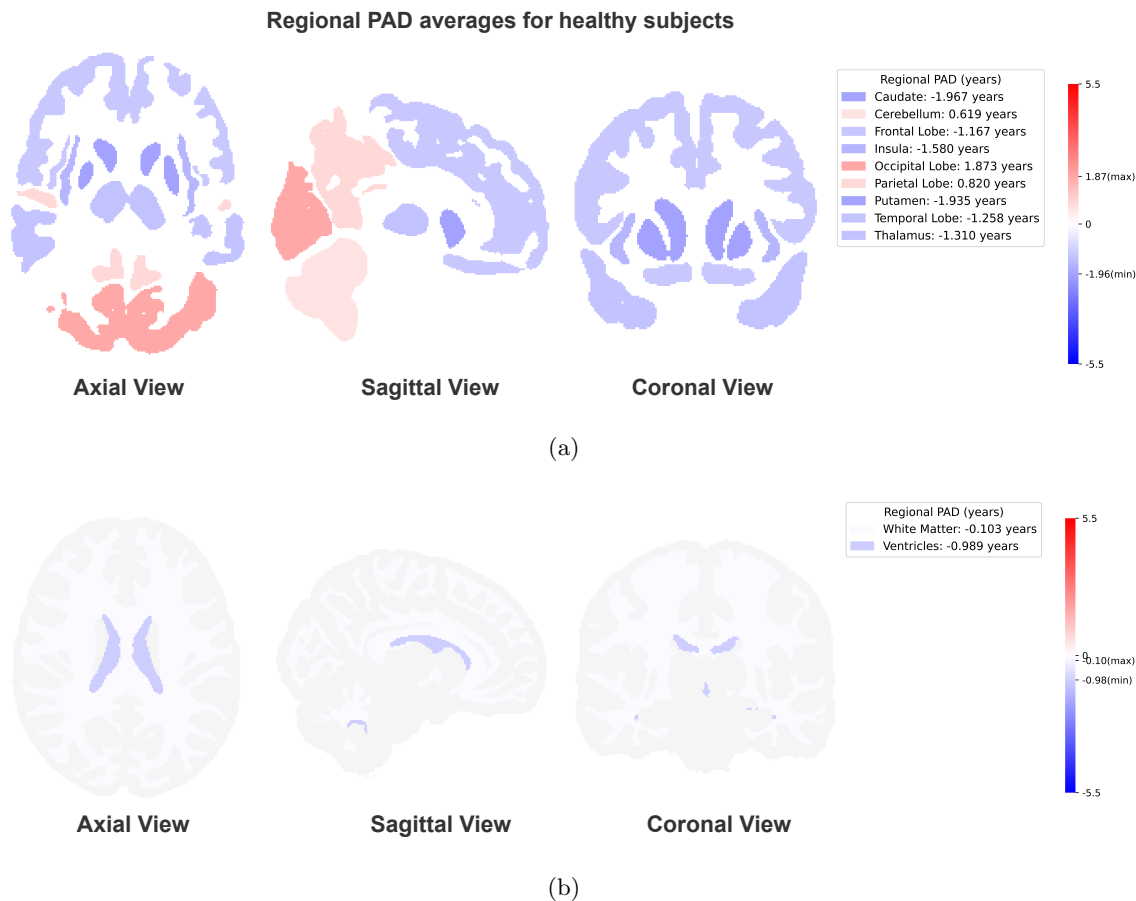


Figure 8: Regional PAD averages for different regions of the brain in a population of presumed healthy subjects. The visualization is in MNI space and has been created using 40 unseen subjects during training from the Cam-CAN dataset. Bias correction as well as filtering for subjects ≤ 70 years of age is done to match the AD and dementia test sets used for regional analysis. It can be observed that most regions of the brain show small negative PAD values with the Caudate looking the youngest with a -1.96 years PAD. Three regions showed positive PAD with Occipital Lobe exhibiting the largest PAD at 1.87 years. Average regional age of the ventricles was -0.98 years making them appear younger than usual.

In Table 4, the regional PAD average and standard deviation (S.D.) values based on the MNI structural atlas (refer to section 3.8) are reported. The regional analysis on three test sets, one corresponding to healthy subjects (Cam-CAN) and two diseased test sets (AD and dementia) was performed. For each dataset, the average (Mean \pm S.D.) PAD values for each region across the test set samples were reported. Additionally, S.D. per region is described (Mean of S.D. \pm S.D. of S.D.) to observe the variability of PAD values within independent regions. To check for statistically significant differences between test sets for

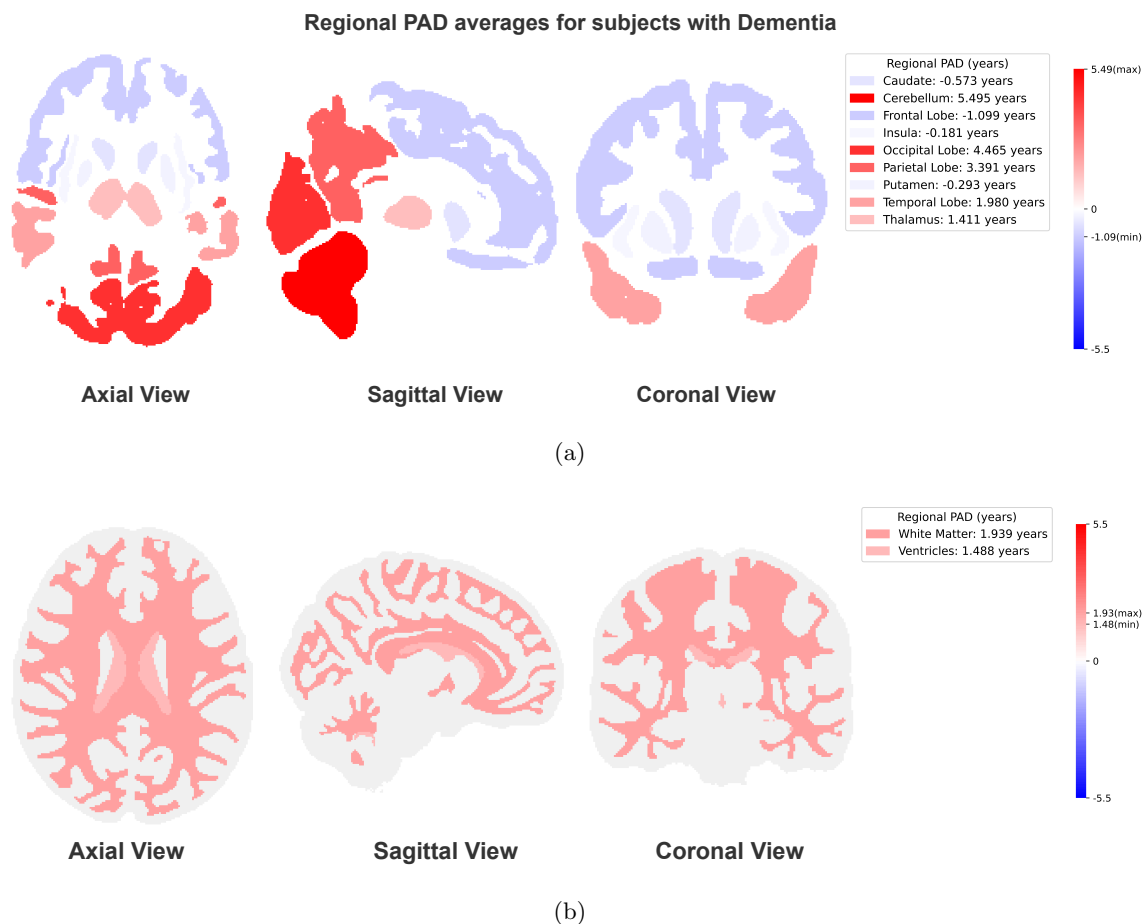


Figure 9: Regional PAD averages for different regions of the brain in a population of subjects with dementia from the OASIS dataset. The visualization is constructed with bias-corrected PAD maps. The visualization is in MNI space and has been created using 15 subjects with age ≤ 70 years using the voxel-level bias-corrected PAD maps. A variation is observed in terms of PAD values across different regions with the Cerebellum, Occipital Lobe, Parietal Lobe, Temporal Lobe, and Thalamus showing an increased brain age (largest PAD of 5.4 years for Cerebellum. Ventricles show an increased brain age with an average regional PAD of 1.48 years.)

each regional average PAD, the Kruskal-Wallis test for performed. This non-parametric test was chosen as the distribution of values was not found to be normal across all populations. Following the Kruskal-Wallis test, Dunn’s test was performed for pairwise comparisons between the possible test set pairs, and multiple comparisons were accounted for using the Holm-Bonferroni correction. For each region, if a significant difference was observed with $p < 0.05$, a * is shown beside the average regional PAD for the test set pair for each region (Table 4).

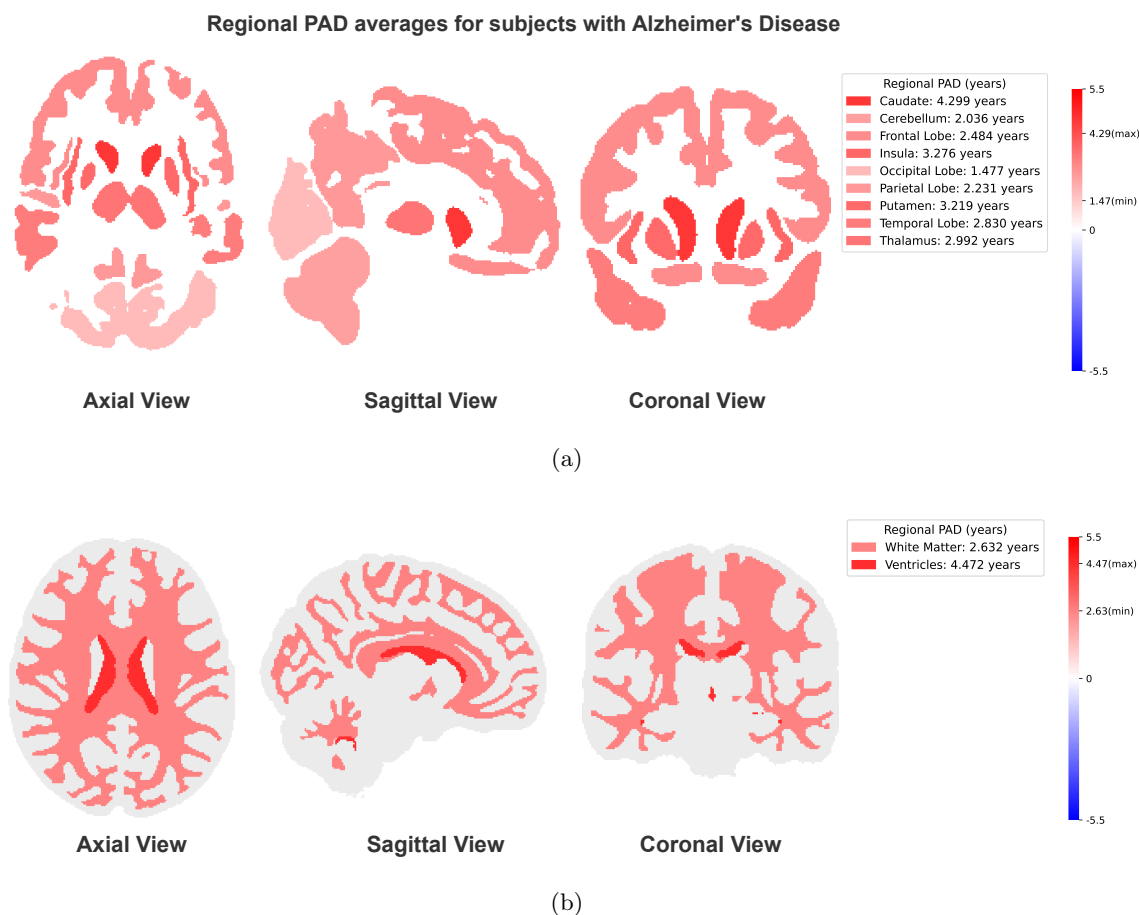


Figure 10: Regional PAD averages for different regions of the brain in a population of subjects with AD from the ADNI dataset. The visualization is constructed with bias-corrected PAD maps. The visualization is in MNI space and has been created using 17 subjects with age ≤ 70 years using the voxel-level bias-corrected PAD maps. It can be observed that all regions in the atlas show an increased brain age with the largest PAD observed in the Ventricles at 4.47 years closely followed by Caudate at 4.29 years.

Figure 8 and Figures 9 and 10 show regional averages of PAD values on the healthy and diseased test sets respectively. A clear distinction can be observed between the healthy versus diseased population averages with the healthy averages appearing to have regional PAD values closer to 0, indicating only a small deviation (less than 2 years) from the chronological age of the subjects. In the averages for diseased subjects (Figures 9 and 10), red colors are observed in most regions of the brain. Overall, the regional averages for subjects with either dementia or AD display an accelerated aging trajectory as well as sharper contrasts as compared to the regional averages corresponding to healthy subjects. It must also be noted that the ventricular region, which is often enlarged as a result of abnormal aging in humans,

appears to show an accelerated aging trajectory for both the dementia and AD populations. To account for the differences in the orientation of different samples in a population, the PAD maps were registered to the MNI space to compute the regional average visualizations as shown in Figures 8 to 10.

Table 4: Regional PAD values. The analysis is done using bias-corrected voxel-level PAD maps for all test sets for consistency.

Regions	Test sets					
	Healthy		AD		Dementia	
	Avg regional PAD	Regional S.D.	Avg regional PAD	Regional S.D.	Avg regional PAD	Regional S.D.
Caudate	$-1.66 \pm 7.85^*$	1.27 ± 0.63	$4.28 \pm 4.49^*$	1.65 ± 0.56	-0.58 ± 11.23	1.39 ± 0.45
Cerebellum	0.38 ± 9.65	3.61 ± 1.64	2.10 ± 5.54	3.60 ± 1.37	5.44 ± 10.80	3.42 ± 1.04
Frontal Lobe	$-1.15 \pm 7.04^*$	2.78 ± 0.91	$2.40 \pm 3.92^*$	3.12 ± 0.69	-1.14 ± 9.55	3.61 ± 1.47
Insula	$-1.56 \pm 8.05^*$	1.70 ± 0.94	$3.29 \pm 3.97^*$	1.54 ± 0.51	-0.21 ± 11.14	1.95 ± 0.87
Occipital Lobe	1.46 ± 8.03	2.89 ± 1.70	1.49 ± 6.13	2.53 ± 0.87	4.37 ± 11.83	2.40 ± 0.90
Parietal Lobe	0.54 ± 7.56	2.61 ± 1.07	2.16 ± 5.15	3.11 ± 0.72	3.31 ± 10.69	3.19 ± 1.36
Putamen	$-1.90 \pm 8.17^*$	1.22 ± 0.57	$3.26 \pm 3.91^*$	1.16 ± 0.38	-0.31 ± 11.20	1.19 ± 0.44
Temporal Lobe	$-1.09 \pm 7.52^*$	3.87 ± 1.80	$2.82 \pm 3.12^*$	3.07 ± 1.00	1.96 ± 9.90	3.72 ± 1.43
Thalamus	$-1.28 \pm 8.46^*$	1.05 ± 0.56	$2.99 \pm 3.69^*$	1.06 ± 0.55	1.41 ± 11.10	1.01 ± 0.32

Note: Statistically significant difference (with $p < 0.05$) between test sets for each region is shown with a *. For example, for Caudate, there is a statistically significant difference between avg. regional PAD for Healthy and AD test sets.

Contribution 4: Interpretability analysis and comparison with traditional interpretability methods: PAD maps obtained from the voxel-level brain age prediction model are compared to the heatmaps obtained from 3 interpretability methods. It is imperative to note that for the scope of this article, the objective of this research is not to propose a state-of-the-art global age prediction model to obtain interpretability maps using traditional methods, however, the aim is to observe the difference in underlying properties and insights obtained from PAD maps versus traditional interpretability heatmaps.

In Figure 11, the first column shows Grad-CAM heatmaps that illustrate regions with relative contribution/importance to the brain age prediction. It is often visualized using red-yellow-blue heatmaps with red regions as the most important and blue being the least. However, since Grad-CAM heatmaps are obtained from the later convolutional layers in a model to observe the final features learned through the gradient with respect to input, they are originally obtained at a much smaller size as compared to input and have to be upsampled, which leads to interpolation errors and coarse maps. The second column shows occlusion sensitivity maps where red regions make the model overestimate the brain age prediction and blue ones make the model underestimate the predictions. White regions

contribute the least. SmoothGrad maps are similar to Grad-CAM heatmaps, except they are generated as a result of multiple forward passes of noisy input through the model to obtain heatmaps that are more precise, counteracting the influence of noise. However, similar to Grad-CAM they are based on the gradients with respect to an input and hence, illustrate the relative importance of regions in one input and are not comparable across samples. Traditional interpretability heatmaps (Grad-CAM, Occlusion Sensitivity maps, and SmoothGrad) maps were originally obtained in the image space registered with 6 degrees of freedom to MNI space, however, for better visualization, the images have been reverted back into the original image space to match the PAD maps in Figure 11.

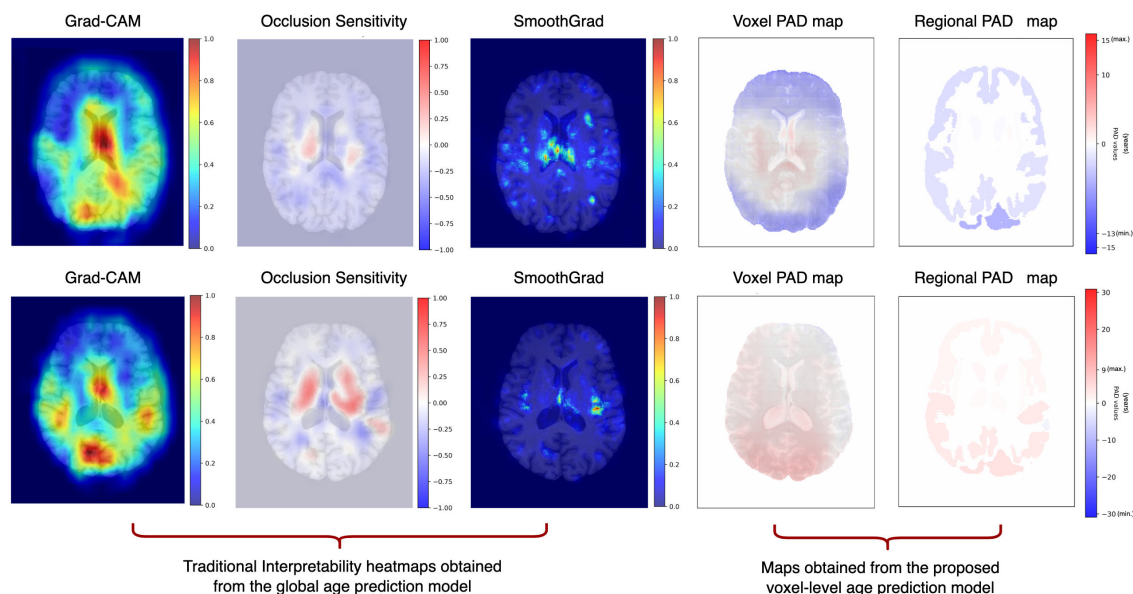


Figure 11: Comparison of traditional interpretability heatmaps (left to right: Grad-CAM, Occlusion Sensitivity, and SmoothGrad) with PAD maps (left to right: voxel-level and regional) obtained from the proposed voxel-level brain age prediction model on presumed healthy subjects. The PAD maps (columns 4 and 5) are non bias-corrected. Grad-CAM and SmoothGrad heatmaps are normalized between 0 to 1. Occlusion Sensitivity maps are bidirectional and hence, normalized between -1 to +1. PAD maps are plotted based on actual PAD values in years.

Voxel-level PAD maps show the regions with an increased brain age in red and decreased brain age in blue. The maps were obtained at the same resolution as the input image due to the upsampling in the U-Net architecture. The use of skip connections in the U-Net architecture leads to accurate upsampling at a high resolution. The intensity values in the PAD maps are quantified in years by computing the difference between predicted and chronological age, and hence, are comparable across samples. The last column in the figure shows the regional PAD maps (PAD values averaged within different known anatomical regions of the brain), which essentially have similar features and characteristics as the voxel-level PAD maps with the difference being in the granularity of the PAD values. This

representation, however, is better suited to analyze the results from the voxel-level age prediction model from an aging perspective.

5. Discussion

The proposed voxel-level brain age prediction model outperforms the baseline U-net voxel-wise prediction model on two independent test sets achieving an error reduction of greater than 30% on both while having a simple and straightforward preprocessing pipeline. Diverging from the baseline (Popescu et al., 2021), the proposed methodology, initially introduced in our previous work Gianchandani et al. (2023), presents two significant modifications. First, the baseline uses non-linear registration as a pre-processing step, registering all T1-weighted images to the MNI template, an average atlas representative of a healthy brain. We hypothesized that each brain structure is unique in terms of shape, size, and structural features and the uniqueness is crucial for brain age estimation. Non-linear registration can modify the uniqueness that each brain volume holds and information is lost in the process. Linear registration can also introduce small smoothening effects to the features of the MR image during translational and rotational changes. Following the same, non-registered images are used as input to the proposed model. This helps retain the original shape, size, and structural features in the truest form possible to be used to predict voxel-level brain age. Second, the baseline uses GM and WM masks obtained from the non-linearly registered images as input to the model, *i.e.*, whole T1-weighted volumes are not fed into the network. The authors rely on implicit CSF features embedded into WM and GM boundaries without actually incorporating CSF segmentation masks. Previous research has shown the relevance of CSF in studying the brain aging process (Houston, 2023; May et al., 1990). Yamada et al. (2023) have shown the increase in intracranial CSF volume as a result of brain volume reduction with age in healthy subjects as well increase in ventricular CSF volume after 60 years of age. Composition changes in CSF volume have also been shown to be associated with AD as well as healthy aging (Fjell et al., 2010). This indicates that CSF volume and composition changes are important indicators to consider when predicting brain age. Hence, instead of relying on the implicitly embedded CSF features in GM and WM masks, the proposed methodology utilizes skull-stripped T1-weighted volumes that include GM, WM, and CSF as input to the model. Segmentation of GM, WM, and CSF is added as one of the output tasks to the proposed model which also contributes to the interpretability analysis. The experiments in this work are designed to present an extended evaluation of our recently proposed voxel-level brain age prediction model using a multi-task approach (Gianchandani et al., 2023). This was done through an evaluation on subjects with underlying neurological disorders, a regional analysis of voxel-level brain age predictions, and an interpretability analysis.

To ensure the learning of accurate feature representations, a subtle noise component (uniform noise between -2 and +2) is introduced to the ground truth labels during model training (refer to Section 3.4). This strategic addition of noise serves to facilitate the model’s ability to discern and understand variations in aging patterns across different brain regions. While this approach introduces noise at the voxel level, it is important to acknowledge that in certain instances, this technique could theoretically yield drastic differences in PAD values between adjacent voxels. For instance, the inclusion of noise might lead to stark contrasts,

such as a red voxel (increased brain age) right adjacent to a contrasting blue voxel (reduced brain age) making the PAD mask appear with a salt and pepper noise appearance. Despite the possibility of sharp contrasts, the PAD maps consistently reveal a tendency toward producing smooth transitions in the brain PAD values with clusters of voxel exhibiting similar patterns of aging. This phenomenon aligns with the inherent nature of aging-related changes, which tend to present on a regional level. Even though the proposed model with intentionally introduced noise performs better than the no-noise version in terms of MAE, this observation in the PAD maps confirms the inclusion of noise does not pose a hindrance or concern in the proposed methodology.

The proposed model’s performance on the sex-stratified CC359 subsets as reported in the results section shows that sex is an important confounder to be considered in brain age prediction studies. The aging process and related structural changes in GM, WM, and CSF vary across different sexes (Gur et al., 1999; Wang et al., 2019) and is reflected in the difference in MAEs on the two sex-specific test sets. To minimize any bias caused by aging differences across males and females, the proposed methodology utilized nearly balanced datasets (for training as well as testing) to ensure equal representation of both sexes. However, future work can include attempts to model voxel-level brain age for sex-specific populations to study the influence of sex in more detail.

The proposed model produces voxel-level PAD maps, which are compared to the heatmaps obtained from traditional interpretability methods. An important feature of the proposed approach that contributes towards ensuring that the proposed model is learning correct features from the input image is the addition of the brain tissue segmentation task as one of the outputs in the architecture. Owing to the multitasking design, the model re-uses the features for the segmentation as well as brain age prediction task. The segmentation performance of the proposed model reached a dice score of 0.89 and 0.83 on the Cam-CAN and CC359 datasets respectively, indicating substantial overlap between predictions and ground truth segmentations. A considerable performance on the segmentation task goes to show that the model learns the structural intricacies within the brain volume such as thickness and shape among other features which are then repurposed for the voxel-level brain age prediction task. This confirms that structural features like changes in GM, WM, and CSF volume, thickness, shape etc. drive model predictions rather than extraneous noise in the background. The inclusion of background noise or any irrelevant features being learned by the proposed model would impact the segmentation performance negatively and be reflected in the dice score, however, that is not the case as the model achieves a considerable dice score on both the internal and external test sets.

Contrary to the heatmaps obtained from traditional interpretability methods which are based on gradients with respect to an input (Grad-CAM, SmoothGrad), the voxel-level PAD maps reflect differences in the prediction from the chronological age in years, making them quantitative and comparable across samples. The occlusion sensitivity maps come close to voxel-PAD maps, however, they are generated by occluding a single region at a time and evaluating its impact on the global age prediction. It is vital to acknowledge that in most machine learning models, multiple regions, which might not adhere to square or cuboid structures, collectively influence final predictions, thus, assessing these regions in isolation is informative, but does not provide the most accurate insight into the collective contributions to brain age predictions. PAD maps, on the other hand, utilize structural

features within the brain region and reflect on voxel-level brain age instead of global brain age, and the results show that the spatial differences in the aging process observed make clinical sense when compared against the structural changes in corresponding T1-weighted images (Gianchandani et al., 2023).

The regional analysis of the PAD maps corresponding to presumed healthy subjects shows PAD values in the narrow range of -1.96 years to 1.87 years (Figure 8), *i.e.*, making most regions (except three) appear slightly younger than the expected chronological age, however, the difference is minimal and can be possibly accounted for by the modeling error. The average regional PAD are closely aligned near 0 (brain age = chronological age), which is the ideal and theoretical scenario, however, does not account for the spatial variations observed in the brain ages across different regions and different samples. It must also be noted that the S.D. (Table 4 - column Average regional PAD) is considerably large for healthy subjects which points to a significant subject-to-subject variability. However, it must be kept in mind that this analysis pertains to a population level encompassing subjects with a diverse age range and unique trajectories of brain aging, all of which might not be reflective in population-level regional averages. Keeping in mind that the aging of the brain is unique to each individual, a subject-level analysis will provide better estimations of regional aging trajectories for the individual.

For the regional analysis, subjects with age ≤ 70 are filtered for the test set. There are two reasons for doing so: (i) The proposed model is trained on subjects up to 88 years of age and to maintain the reliability of predictions, a deliberate choice was made to refrain from evaluating the model on subjects exceeding 88 years of age. The predictions in the peripheral regions of the in-domain Cam-CAN test set (ages 70 and above as shown in Figure 3) are often observed to exhibit a bias, leading to under-prediction or younger-looking brains for older age ranges. While the bias is addressed through a dedicated correction process as explained in Section 3.7, it is important to note that the methodology used for this bias correction is built upon data from healthy subjects. It is tailored to the patterns observed in the evaluation of healthy subjects. It would be unfair to assume that, for diseased subjects, the same bias correction methodology would suffice to mitigate the bias observed. (ii) Based on the bias-correction methodology, a different correcting factor is used for different age ranges and theoretically, if diseased subjects are expected to have an increased brain age relative to the corresponding chronological age, it would be unfair to use the correcting factor based on the chronological age as the bias observed would be relative to an older age (compared to the chronological age). Hence, to ensure that bias correction does not fail significantly, and helps with mitigating the bias to a reasonable extent, this precautionary filtering is performed to remove subjects with age ≥ 71 years. Nonetheless, since most neurological disorders are observed in an older population, bias correction becomes imperative for the AD and dementia test sets for the regional analysis to help account for the bias, even though it might not mitigate the bias entirely. For consistency of results, precautionary filtering and bias-correction were also performed for the healthy test set for regional analysis.

Another important consideration when analyzing the regional PAD values in Table 4 is that in the case of a healthy population, the age range of subjects is wide enough (18-70 years) such that the small bias observed is in both directions as over-predictions and as well as under-predictions. Hence, at the population level, the over and under-predictions tend

to cancel each other's effect to an extent. However, this might not be the case for diseased subjects as most subjects in the test set are above the average training set age and hence, bias is only observed in the form of under-predictions (*i.e.* negative PAD). As mentioned previously, bias-correction does not account for 100% of the bias in diseased subjects coupled with the fact that only under-predictions are observed, the results of the PAD values in Table 4 and Figures 9 and 10 might still reflect a small degree of bias and be more negative than the actual values.

The regional PAD values, MNI atlases, and PAD maps corresponding to individual subjects were reviewed by a radiologist (JO) and some notable observations were made:

1. It can be observed that in subjects with dementia, ventricles tend to show an accelerated brain age as compared to the rest of the brain regions (refer to adjusted PAD maps in Figure 7). It is unclear whether this increased aging of the ventricles is mostly related to an increase in ventricle size, which is usually a sequelae of generalized brain parenchymal volume loss, or due to differences in CSF composition. Both these explanations seem plausible: large ventricle size is associated with the presence of neurodegenerative disorders, and even in healthy subjects, increased ventricle volume seems to indicate a greater risk of developing dementia in the future (Carmichael et al., 2007). Furthermore, cellular CSF composition is altered in subjects with neurodegenerative diseases, with a shift from central memory to effector T cells (Busse et al., 2021). Such changes do not affect MR image signal intensity in any noticeable way upon visual inspection by radiologists, but there may be subtle signal changes that may have been detected by the proposed model.
2. In AD subjects, PAD was particularly high in the Caudate nuclei with an average of ~ 4.29 years (Figure 10). Previous studies have found lower Caudate nuclei volumes in AD compared to healthy control subjects (Madsen et al., 2010). Considering the possible association between lower Caudate volume and advanced brain age, these prior findings are in line with the results of the current study. It is important to emphasize that this is a speculative interpretation of the results presented in this work and should not be translated to clinical applications without further analysis. Increased brain age (2.8 years) was also observed in the Temporal Lobe with a high regional standard deviation indicating a great degree of variation within the region, which is often an important region associated with AD.
3. In the group of dementia subjects, brain age was particularly advanced in the posterior brain regions, *i.e.*, the Occipital and posterior Parietal lobes, and the Cerebellum (Figure 9). Atrophy predominantly affecting the posterior brain parenchyma is uncommon in dementia patients. It can sometimes be seen in AD patients (Crutch et al., 2012) and is a hallmark feature of Lewy body dementia, a rare neurodegenerative disease (Silva-Rodríguez et al., 2023). The exact underlying dementia etiologies are not known in the dementia subgroup of this study; there is a possibility that some of these subjects did suffer from Lewy body dementia or posterior predominant AD. While previous studies mainly focused on brain parenchymal volume, the proposed model predicts brain age using a multidimensional approach. It is possible that characteristics other than volume, for example, changes in brain signal intensity or structure, occur in subjects with dementia that do not affect volume and are, therefore, not well known yet. This, however, is speculative since we do not have exact clinical diagnoses and therefore can neither confirm nor refute this hypothesis.
4. In the group of dementia subjects, negative PAD was observed in certain regions like the

Caudate, Frontal Lobe, Insula, and Putamen. Since the exact underlying dementia etiologies are not known in this group, the negative values could be a grouped effect of multiple different types of dementia observed (or not observed) in the subjects of this group. This group showed increased brain age in the posterior regions of the brain which points to the possible presence of subjects with posterior predominant dementia, whereas a negative PAD for the Frontal Lobe indicates the absence of subjects with frontal predominant dementia. Insular atrophy or degeneration is also observed in the presence of frontotemporal dementia (Seeley, 2010) and could be the driving factor for a negative PAD. Putamen is dominantly associated with the presence of Parkinson’s disease (Kinoshita et al., 2022), however, it averaged at -0.31 years which is a rather small deviation from zero and could simply be an impact of modeling error. It must also be noted that a high standard deviation (Table 4: Avg Regional PAD column) is observed for the dementia test set as compared to others that indicates subject-to-subject variation, which is plausible as this test set includes subjects with different dementia types and hence, differences in regional brain aging trajectories.

The findings from the the proposed brain age prediction model are partially consistent with the known biomarkers of aging in subjects with dementia and more specifically, AD. Some new potential biomarkers like increased brain age in posterior regions of the brain have been identified by the proposed model, and require further validation.

It is crucial to emphasize that though it is important to understand regional aging patterns for older subjects *i.e.*, where disorders are observed and are often progressed to a stage where the subject exhibits noticeable symptoms and is already a part of the research study collecting data; another important aim of this research is to predict early onset of neurological disorders before the subjects start exhibiting symptoms and apparent cognitive decline. Therefore, evaluation on healthy subjects is an important part as it can unveil potential indicators of early onset of neurological disorders. A future direction to validate the proposed model would be to evaluate the model on longitudinal data which includes subjects transitioning from an initial presumed healthy stage to some form of underlying neurological disorder.

6. Conclusion

In this study, previous analysis of a voxel-level brain age prediction model is extended as a proof-of-concept. Through the experiments, the choice of a multitask architecture is validated and it is shown that using a voxel-level approach can be a way of achieving improved interpretability and a better understanding of regional aging trajectories. Evaluation of the model on healthy subjects as well as ones with dementia and specifically, AD revealed consistent findings on regional brain aging as other aging studies and also revealed new indicators that can be potential biomarkers of the presence of dementia. Through this research, the transition of brain age prediction models towards voxel-level predictions is shown as a way to enhance the understanding of the degenerating brain while demonstrating an improvement with respect to existing implementations.

Acknowledgments

NG is supported by the Natural Sciences and Engineering Research Council (NSERC) BRAIN CREATE award and the Alberta Innovates Graduate Student Scholarship. RS thanks the NSERC (RGPIN/2021-02867) for ongoing operating support for this project. RS also thanks the Hotchkiss Brain Institute for financial support. MEM acknowledges support from startup funding at the University of Calgary and the NSERC Discovery Grant (RGPIN-03552) and Early Career Researcher Supplement (DGEGR-00124). Data collection and sharing for this project was partly funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript. All data used in this study was obtained from publicly available datasets and has been handled following the terms provided by the data sources. Data anonymity has been maintained and all data sources have been properly cited complying with ethical and privacy regulations.

Conflicts of Interest

The authors have no competing interests to declare.

References

- Habtamu M Aycheh, Joon-Kyung Seong, Jeong-Hyeon Shin, Duk L Na, Byungkon Kang, Sang W Seo, and Kyung-Ah Sohn. Biological brain age prediction using cortical thickness data: a large scale cohort study. *Frontiers in aging neuroscience*, 10:252, 2018.
- Gareth Ball, Claire E Kelly, Richard Beare, and Marc L Seal. Individual variation underlying brain age estimates in typical development. *Neuroimage*, 235:118036, 2021.
- Iman Beheshti, Pierre Gravel, Olivier Potvin, Louis Dieumegarde, and Simon Duchesne. A novel patch-based procedure for estimating brain age across adulthood. *Neuroimage*, 197: 618–624, 2019.
- Iman Beheshti, MA Ganaie, Vardhan Paliwal, Aryan Rastogi, Imran Razzak, and Muhammad Tanveer. Predicting brain age using machine learning algorithms: A comprehensive evaluation. *IEEE Journal of Biomedical and Health Informatics*, 26(4):1432–1440, 2021.
- Camilo Bermudez, Andrew J Plassard, Shikha Chaganti, Yuankai Huo, Katherine S Aboud, Laurie E Cutting, Susan M Resnick, and Bennett A Landman. Anatomical context improves deep learning on the brain age estimation task. *Magnetic Resonance Imaging*, 62:70–77, 2019.
- Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain

- mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86: 102789, 2023.
- Kyriaki-Margarita Bintsi, Vasileios Baltatzis, Arinbjörn Kolbeinsson, Alexander Hammers, and Daniel Rueckert. Patch-based brain age estimation from MR images. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, pages 98–107. Springer, 2020.
- Kyriaki-Margarita Bintsi, Vasileios Baltatzis, Alexander Hammers, and Daniel Rueckert. 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 4*, pages 65–74. Springer, 2021.
- Fabio Blandini, Giuseppe Nappi, Cristina Tassorelli, and Emilia Martignoni. Functional changes of the basal ganglia circuitry in parkinson’s disease. *Progress in neurobiology*, 62(1):63–88, 2000.
- Stefan Busse, Jessica Hoffmann, Enrico Michler, Roland Hartig, Thomas Frodl, and Mandy Busse. Dementia-associated changes of immune cell composition within the cerebrospinal fluid. *Brain, Behavior, & Immunity-Health*, 14:100218, 2021.
- Daniele Caligiore, Rick C Helmich, Mark Hallett, Ahmed A Moustafa, Lars Timmermann, Ivan Toni, and Gianluca Baldassarre. Parkinson’s disease as a system-level disorder. *npj Parkinson’s Disease*, 2(1):1–9, 2016.
- M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.
- Owen T Carmichael, Lewis H Kuller, Oscar L Lopez, Paul M Thompson, Rebecca A Dutton, Allen Lu, Sharon E Lee, Jessica Y Lee, Howard J Aizenstein, Carolyn Cidis Meltzer, et al. Ventricular volume and dementia progression in the cardiovascular health study. *Neurobiology of aging*, 28(3):389–397, 2007.
- James H Cole. Neuroimaging-derived brain-age: an ageing biomarker? *Aging (Albany NY)*, 9(8):1861, 2017.
- James H Cole and Katja Franke. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in neurosciences*, 40(12):681–690, 2017.
- James H Cole, Rudra PK Poudel, Dimosthenis Tsagkrasoulis, Matthan WA Caan, Claire Steves, Tim D Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163: 115–124, 2017.
- James H Cole, Stuart J Ritchie, Mark E Bastin, Valdés Hernández, S Muñoz Maniega, Natalie Royle, Janie Corley, Alison Pattie, Sarah E Harris, Qian Zhang, et al. Brain age predicts mortality. *Molecular psychiatry*, 23(5):1385–1392, 2018.

- D Louis Collins, Colin J Holmes, Terrence M Peters, and Alan C Evans. Automatic 3-d model-based neuroanatomical segmentation. *Human brain mapping*, 3(3):190–208, 1995.
- Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020.
- Sebastian J Crutch, Manja Lehmann, Jonathan M Schott, Gil D Rabinovici, Martin N Rossor, and Nick C Fox. Posterior cortical atrophy. *The Lancet Neurology*, 11(2):170–178, 2012.
- Richard J Davidson, Heather Abercrombie, Jack B Nitschke, and Katherine Putnam. Regional brain function, emotion and disorders of emotion. *Current opinion in neurobiology*, 9(2):228–234, 1999.
- Xinyang Feng, Zachary C Lipton, Jie Yang, Scott A Small, Frank A Provenzano, Alzheimer’s Disease Neuroimaging Initiative, Frontotemporal Lobar Degeneration Neuroimaging Initiative, et al. Estimating brain age based on a uniform healthy population with deep learning and structural magnetic resonance imaging. *Neurobiology of aging*, 91:15–25, 2020.
- Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.
- Anders M Fjell, Kristine B Walhovd, Christine Fennema-Notestine, Linda K McEvoy, Donald J Hagler, Dominic Holland, Kaj Blennow, James B Brewer, Anders M Dale, and Alzheimer’s Disease Neuroimaging Initiative. Brain atrophy in healthy aging is related to csf levels of $a\beta$ 1-42. *Cerebral Cortex*, 20(9):2069–2079, 2010.
- Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, D Louis Collins, Brain Development Cooperative Group, et al. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313–327, 2011.
- Katja Franke and Christian Gaser. Ten years of brainage as a neuroimaging biomarker of brain aging: what insights have we gained? *Frontiers in neurology*, page 789, 2019.
- Neha Gianchandani, Johanna Ospel, Ethan MacDonald, and Roberto Souza. A multitask deep learning model for voxel-level brain age estimation. In *International Workshop on Machine Learning in Medical Imaging*. Springer, 2023. accepted for publication.
- Weikang Gong, Christian F Beckmann, Andrea Vedaldi, Stephen M Smith, and Han Peng. Optimising a simple fully convolutional network for accurate brain age prediction in the PAC 2019 challenge. *Frontiers in Psychiatry*, 12:627996, 2021.
- Ruben C Gur, Bruce I Turetsky, Mie Matsui, Michelle Yan, Warren Bilker, Paul Hughett, and Raquel E Gur. Sex differences in brain gray and white matter in healthy young adults: correlations with cognitive performance. *Journal of Neuroscience*, 19(10):4065–4072, 1999.
- Sheng He, P Ellen Grant, and Yangming Ou. Global-local transformer for brain age estimation. *IEEE transactions on medical imaging*, 41(1):213–224, 2021.

- PR Hof, Pantaleimon Giannakopoulos, and Constantin Bouras. The neuropathological changes associated with normal brain aging. *Histology and histopathology*, 1996.
- Simon M Hofmann, Frauke Beyer, Sebastian Lapuschkin, Ole Goltermann, Markus Loeffler, Klaus-Robert Müller, Arno Villringer, Wojciech Samek, and A Veronica Witte. Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain. *NeuroImage*, 261:119504, 2022.
- Stephanie Houston. Aging in the csf. *Nature Immunology*, 24(2):203–203, 2023.
- Tzu-Wei Huang, Hwann-Tzong Chen, Ryuichi Fujimoto, Koichi Ito, Kai Wu, Kazunori Sato, Yasuyuki Taki, Hiroshi Fukuda, and Takafumi Aoki. Age estimation from brain mri images using deep learning. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 849–852. IEEE, 2017.
- Koichi Ito, Ryuichi Fujimoto, Tzu-Wei Huang, Hwann-Tzong Chen, Kai Wu, Kazunori Sato, Yasuyuki Taki, Hiroshi Fukuda, and Takafumi Aoki. Performance evaluation of age estimation from T1-weighted images using brain local features and CNN. In *IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 694–697. IEEE, 2018.
- Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. FSL. *Neuroimage*, 62(2):782–790, 2012.
- Keisuke Kinoshita, Takehito Kuge, Yoshie Hara, and Kojiro Mekata. Putamen atrophy is a possible clinical evaluation index for parkinson’s disease using human brain magnetic resonance imaging. *Journal of Imaging*, 8(11):299, 2022.
- Arinbjörn Kolbeinsson, Sarah Filippi, Yannis Panagakis, Paul M Matthews, Paul Elliott, Abbas Dehghan, and Ioanna Tzoulaki. Accelerated MRI-predicted brain ageing and its associations with cardiometabolic and brain disorders. *Scientific Reports*, 10(1):1–9, 2020.
- Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, pages 2019–12, 2019.
- Herve Lemaitre, Aaron L Goldman, Fabio Sambataro, Beth A Verchinski, Andreas Meyer-Lindenberg, Daniel R Weinberger, and Venkata S Mattay. Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume? *Neurobiology of aging*, 33(3):617–e1, 2012.
- Hualou Liang, Fengqing Zhang, and Xin Niu. Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. *Human Brain Mapping*, 40(11):3143, 2019.
- M Ethan MacDonald and G Bruce Pike. MRI of healthy brain aging: A review. *NMR in Biomedicine*, 34(9):e4564, 2021.

- Sarah K Madsen, April J Ho, Xue Hua, Priya S Saharan, Arthur W Toga, Clifford R Jack Jr, Michael W Weiner, Paul M Thompson, Alzheimer's Disease Neuroimaging Initiative, et al. 3d maps localize caudate nucleus atrophy in 400 alzheimer's disease, mild cognitive impairment, and healthy elderly subjects. *Neurobiology of aging*, 31(8):1312–1325, 2010.
- Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. 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, 2007.
- C May, JA Kaye, John R Atack, MB Schapiro, RP Friedland, and SI Rapoport. Cerebrospinal fluid production is reduced in healthy aging. *Neurology*, 40(3 Part 1):500–500, 1990.
- John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322, 2001.
- Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics*, 15(4):869–877, 2005a.
- Susanne G Mueller, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford R Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's & Dementia*, 1(1):55–66, 2005b.
- Lorenzo Pasquini, Farzaneh Rahmani, Somayeh Maleki-Balajoo, Renaud La Joie, Mojtaba Zarei, Christian Sorg, Alexander Drzezga, and Masoud Tahmasian. Medial temporal lobe disconnection and hyperexcitability across alzheimer's disease stages. *Journal of Alzheimer's disease reports*, 3(1):103–112, 2019.
- Han Peng, Weikang Gong, Christian F Beckmann, Andrea Vedaldi, and Stephen M Smith. Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis*, 68:101871, 2021.
- Sergey M Plis, Devon R Hjelm, Ruslan Salakhutdinov, Elena A Allen, Henry J Bockholt, Jeffrey D Long, Hans J Johnson, Jane S Paulsen, Jessica A Turner, and Vince D Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience*, 8:229, 2014.
- Sebastian G Popescu, Ben Glocker, David J Sharp, and James H Cole. Local brain-age: a u-net model. *Frontiers in Aging Neuroscience*, 13:761954, 2021.
- Y Lakshmisha Rao, B Ganaraja, BV Murlimanju, Teresa Joy, Ashwin Krishnamurthy, and Amit Agrawal. Hippocampus and its involvement in Alzheimer's disease: a review. *3 Biotech*, 12(2):55, 2022.

- Naftali Raz, Ulman Lindenberger, Karen M Rodrigue, Kristen M Kennedy, Denise Head, Adrienne Williamson, Cheryl Dahle, Denis Gerstorff, and James D Acker. Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral cortex*, 15(11):1676–1689, 2005.
- Naftali Raz, Paolo Ghisletta, Karen M Rodrigue, Kristen M Kennedy, and Ulman Lindenberger. Trajectories of brain aging in middle-aged and older adults: regional and individual differences. *Neuroimage*, 51(2):501–511, 2010.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- Hedieh Sajedi and Nastaran Pardakhti. Age prediction based on brain mri image: a survey. *Journal of medical systems*, 43:1–30, 2019.
- Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in neural information processing systems*, 31, 2018.
- William W Seeley. Anterior insula degeneration in frontotemporal dementia. *Brain Structure and Function*, 214:465–475, 2010.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE ICCV*, pages 618–626, 2017.
- Jesús Silva-Rodríguez, Miguel A Labrador-Espinosa, Alexis Moscoso, Michael Schöll, Pablo Mir, and Michel J Grothe. Characteristics of amnesic patients with hypometabolism patterns suggestive of lewy body pathology. *Brain*, page awad194, 2023.
- Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.
- Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain MR dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 170:482–494, 2018.
- Jason R Taylor, Nitin Williams, Rhodri Cusack, Tibor Auer, Meredith A Shafto, Marie Dixon, Lorraine K Tyler, Richard N Henson, et al. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *neuroimage*, 144:262–269, 2017. .
- SA Valizadeh, Jürgen Hänggi, Susan Mérellat, and Lutz Jäncke. Age prediction on the basis of brain anatomical measures. *Human brain mapping*, 38(2):997–1008, 2017.

- Yanpei Wang, Qinfang Xu, Jie Luo, Mingming Hu, and Chenyi Zuo. Effects of age and sex on subcortical volumes. *Frontiers in aging neuroscience*, 11:259, 2019.
- Shigeki Yamada, Tomohiro Otani, Satoshi Ii, Hiroto Kawano, Kazuhiko Nozaki, Shigeo Wada, Marie Oshima, and Yoshiyuki Watanabe. Aging-related volume changes in the brain and cerebrospinal fluid using artificial intelligence-automated segmentation. *European Radiology*, pages 1–14, 2023.
- Chenzhong Yin, Phoebe Imms, Mingxi Cheng, Anar Amgalan, Nahian F Chowdhury, Roy J Massett, Nikhil N Chaudhari, Xinghe Chen, Paul M Thompson, Paul Bogdan, et al. Anatomically interpretable deep learning of brain age captures domain-specific cognitive impairment. *Proceedings of the National Academy of Sciences*, 120(2):e2214634120, 2023.
- Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.