Received: 11 November 2019 | Revised: 27 February 2020 | Accepted: 7 April 2020

DOI: 10.1002/hbm.25011

RESEARCH ARTICLE

WILEY

From a deep learning model back to the brain—Identifying regional predictors and their relation to aging

Gidon Levakov[1,2] | Gideon Rosenthal[1,2] | Ilan Shelef[2,3] | Tammy Riklin Raviv[2,4] | Galia Avidan[1,2,5]

[1]Department of Cognitive and Brain Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[2]Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[3]Department of Diagnostic Imaging, Ben-Gurion University of the Negev, Beer-Sheva, Israel

[4]The School of Electrical and Computer Engineering, Ben Gurion University of the Negev, Beer-Sheva, Israel

[5]Department of Psychology, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Correspondence
Gidon Levakov, Department of Cognitive and Brain Sciences, Ben-Gurion University of the Negev, Beer-Sheva, Israel.
Email: gidonle@post.bgu.ac.il

Funding information
Ben Gurion University of the Negev, Grant/Award Number: Internal funding grant

Abstract

We present a Deep Learning framework for the prediction of chronological age from structural magnetic resonance imaging scans. Previous findings associate increased brain age with neurodegenerative diseases and higher mortality rates. However, the importance of brain age prediction goes beyond serving as biomarkers for neurological disorders. Specifically, utilizing convolutional neural network (CNN) analysis to identify brain regions contributing to the prediction can shed light on the complex multivariate process of brain aging. Previous work examined methods to attribute pixel/voxel-wise contributions to the prediction in a single image, resulting in "explanation maps" that were found noisy and unreliable. To address this problem, we developed an inference scheme for combining these maps across subjects, thus creating a population-based, rather than a subject-specific map. We applied this method to a CNN ensemble trained on predicting subjects' age from raw T1 brain images in a lifespan sample of 10,176 subjects. Evaluating the model on an untouched test set resulted in mean absolute error of 3.07 years and a correlation between chronological and predicted age of $r = 0.98$. Using the inference method, we revealed that cavities containing cerebrospinal fluid, previously found as general atrophy markers, had the highest contribution for age prediction. Comparing maps derived from different models within the ensemble allowed to assess differences and similarities in brain regions utilized by the model. We showed that this method substantially increased the replicability of explanation maps, converged with results from voxel-based morphometry age studies and highlighted brain regions whose volumetric variability correlated the most with the prediction error.

KEYWORDS
brain aging, convolutional neural networks, deep learning, interpretability, neuroimaging

1 | INTRODUCTION

The human brain undergoes complex structural changes across the lifespan (Sowell, Thompson, & Toga, 2004). These include widespread synaptic pruning and myelination from early life through puberty and neurodegenerative processes, such as ventricle expansion and cortical

Tammy Riklin Raviv and Galia Avidan have contributed equally to this work.

Data used in preparation of this article were partially obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu) and the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO). As such, the investigators within ADNI and AIBL contributed to the design and implementation of ADNI and AIBL and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI and AIBL investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf and at www.aibl.csiro.au.

thinning that peaks with aging. The course and extent of these changes are not uniformly distributed across the brain (Storsve et al., 2014). Thus, for example in healthy aging, higher atrophy rates were reported in the hippocampus, while regions like the early visual cortex remain relatively intact (but see Lemaitre et al., 2012). Nevertheless, studies that examined the correspondence between brain structure and chronological age provide inconsistent findings. Such inconsistencies may be related to the specific parcellation schemes employed (Mikhael & Pernet, 2019), surface-based structural measurements (Lemaitre et al., 2012), global volume covariates (Jäncke, Mérillat, Liem, & Hänggi, 2015), or specific statistical assumptions regarding the changes of brain-aging rate across lifespan (e.g., linear, polynomial; Ziegler et al., 2012). These concerns add to discrepancies due to the usage of relatively small samples and different statistical procedures, which together impede the attempts to characterize the relation between aging and structural changes in the brain.

Studying brain aging has important implications for differentiating typical and pathological aging. Alzheimer's disease (AD), the most prevalent type of dementia, affects about 22% of the population over the age of 75 (in the United States, 2010; Hebert, Weuve, Scherr, & Evans, 2013). AD patients exhibit extensive cell loss in cortical and subcortical regions, but such findings are also evident in typical aging (Barnes et al., 2009; Ledig, Schuh, Guerrero, Heckemann, & Rueckert, 2018). Moreover, behavioral manifestations such as cognitive decline and memory deficits that accompany AD are also apparent in aging in the absence of AD (Cardenas et al., 2011; Koen & Yonelinas, 2014). Thus, a reliable measure of typical brain aging may be beneficial in order to better distinguish between the two (Lorenzi, Pennec, Frisoni, & Ayache, 2015).

1.1 | Predicting age from structural brain imaging using machine learning

Recent growth in data availability and advancements in the field of machine learning (ML), applied to the analysis of structural imaging, have allowed addressing regression problems such as brain age prediction based on preselected sets of anatomical features or regions of interest (ROIs). Predicting age from brain anatomy enables to estimate a measure of "brain age" which is independent of one's chronological age. Different studies generally reveal that an over-estimation of that measure is associated with neurodegenerative diseases and various clinical conditions and might even predict mortality (Cole et al., 2018). Hence, brain age estimation could be used as a potential biomarker for brain health (Cole & Franke, 2017). While ML methods were shown to provide a mean error of ~5 years (Cole & Franke, 2017), age predictions are largely dependent upon the selection of features that would be used as input to the algorithm.

1.2 | Application of deep convolutional neural network for predicting "brain age"

Deep convolutional neural network (CNN) has enabled a major leap in many applications including neuroimaging analysis, among others, by learning the features, or representation from the raw data, that is, an image or a volume (Goodfellow, Bengio, & Courville, 2016). CNNs are biologically inspired algorithms in which the connectivity between the different neurons implements a convolution operation. The neurons are ordered in stacked layers in a hierarchical deep formation and hence they are termed deep CNN (LeCun, Bottou, Bengio, & Haffner, 1998). CNN-based models achieved state-of-the-art results in serval neuroimaging tasks including cortical segmentation and tumor detection (Kamnitsas, Chen, & Ledig, 2015; Pereira, Pinto, Alves, & Silva, 2016) and were recently applied to age prediction from raw T1 magnetic resonance imaging (MRI) images (Cole et al., 2017). Nonetheless, significant improvement can still be achieved by substantially increasing the sample size and utilizing practices such as prediction based on an ensemble of models. Both of these approaches were shown to produce a remarkable improvement in other visual task domains (Lee, Purushwalkam, Cogswell, Crandall, & Batra, 2015).

1.3 | Model interpretability—Which brain regions underlie a given prediction?

A major limitation of studies utilizing CNNs pertains to the issue of the model interpretability. While CNNs have provided high accuracy for age prediction (Cole & Franke, 2017; Qi, Du, Zhuang, Huang, & Ding, 2018), it is typically difficult to identify the features that enabled a given prediction. Several recent studies attempted to identify or visualize intermediate representations of the CNN (Olah et al., 2018), but still, the size and complexity of the networks make it a challenging task. In the context of structural neuroimaging analysis, there might be an advantage to focus on the input level since it could be directly related to specific brain structures. Knowing which image parts, or in the current research, brain regions or neural attributes, contribute most to the prediction have theoretical, as well as translational value. A possible approach to this issue is the usage of "saliency maps" or "explanation maps" indicating the influence of each voxel in the input volume on the model's prediction. Such a map can be generated by calculating the partial derivative of each voxel in the input volume with respect to the model's output (Simonyan, Vedaldi, & Zisserman, 2013; Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014). However, local gradients in nonlinear functions such as CNN were previously shown to be noisy. Recent work has demonstrated that this could be partially addressed by repeatedly calculating and averaging several explanation maps derived from the same input after adding random noise to it (Smilkov, Thorat, Kim, Viégas, & Wattenberg, 2017). Nevertheless, these explanation maps are typically created on a single sample, hence they provide only a subject-specific, rather than a population-based explanation (Yang, Rangarajan, & Ranka, 2018; but see Bermudez et al., 2019; Wang et al., 2019). In a task or a model where large variability exists in these explanation maps, that is, if different subject-level maps highlight different regions, any translational or theoretical conclusion could only be subject-specific.

1.4 | The current study

In light of the limitations outlined above, we aimed to examine brain aging using a CNN model for "brain age" prediction and identify the brain structures that supported this prediction. Therefore, this study has two important contributions. The first is the prediction model, which is composed of an ensemble of multiple CNNs trained to predict individuals' age from minimally processed T1 MRI scans. The model was trained and tested on an aggregated lifespan sample of 10,176 subjects. These were collected from several large-scale open-access databases (*n* = 15) in order to produce a result that is more robust to scanner's type, field strength, and resolution. Second, we provided and validated a novel scheme for identifying the importance of the various anatomical brain regions to the age prediction by aggregating multiple subject-level explanation maps, creating a population-based map. Combining subject-level maps into a population-based map is done by image realignment after training the model, thus no special preprocessing or architecture modification is required, as opposed to previous work (Ito et al., 2018). We empirically show that this significantly improves the explanation maps and allows the inference from the model back to the brain's anatomy. Finally, we demonstrate how the usage of an ensemble of CNNs which increases the prediction accuracy, also allows to evaluate the diversity or similarity of independently trained models, to asses model uncertainty, and to examine the extent to which different models exploit similar brain regions for the age prediction.

2 | MATERIALS AND METHODS

2.1 | Datasets

To train a model that would be more robust to different sites and scanning protocols, we collected a sample of 10,176 T1w MRI brain scans of individuals ranging between 4 and 94 years old from various open databases (*n* = 15), acquired at different locations, scanners, and scanning parameters. To examine whether the trained model presents a dataset-specific bias, that is, over or under-estimated age within a specific set, we examined the mean signed error for each study. The analysis was conducted within the test set and was limited for studies with more than 15 test samples (4 excluded, 11 remained). Several databases from longitudinal studies consist of brain scans acquired at several time-points. For these databases, we only used scans of the first time point to avoid data leakage between the train and validation/test sets. Three exclusion criteria were applied to all subjects: missing age report, major artifacts in a visual inspection of the T1 volume and diagnosis of AD or another form of dementia. The complete list of studies, age, and gender distributions are reported in Table 1.

2.2 | Data preprocessing

To minimize the model reliance on preprocessing steps such as image realignment and registration that are both computationally intensive and time-consuming, we designed a minimal preprocessing procedure. To ensure that the model "brain age" estimation would rely solely on regions within the skull, the only substantial preprocessing step was the removal of extra-cranial regions from the volume. Thus, the preprocessing procedure included four stages: applying a coarse (90°) rotation so that all the volumes would appear in similar L–R, A–P, S–I orientation (FSL *fslreorient2std* tool; Woolrich et al., 2009), skull removing tool (ROBEX; Iglesias, Liu, Thompson, & Zhuowen, 2011), volume resize to standard size (90, 126, 110) and volume standardization (μ = 0, σ = 1). Resizing of each volume was implemented by applying an identical scaling factor to all three dimensions, such that brain voxels (intensity >0) would occupy the maximum portion within the final volume (90, 126, 110). For each volume, voxels' intensities were standardized by a subtraction of the volume's mean intensity followed by division by the intensity's *SD*.

2.3 | Data augmentation

Head orientation, the field of view and the level of signal to noise ratio may differ between scans even if they were acquired by the same machine and are of the same subject. To improve the robustness of the models to these variations, we augmented the training data by randomly manipulating the head position, size, and noise level. This procedure was previously shown to improve generalization and avoid overfitting (Simard, Steinkraus, & Platt, 2003). Specifically, the series of transformation to the brain image included a rotation in the x/y/z-axis *unif*(−10°, 10°), shifting *unif*(−5, 5) voxels, scaling $\mathcal{N}(0,0.1)$ and an addition of random noise $\mathcal{N}(0,0.015)$. Data augmentation was applied only during training and was not used for validation/test. The optimal augmentation parameters were chosen as the ones that maximized the validation accuracy using a random hyperparameter search.

2.4 | CNN architecture

The CNN models were implemented using Keras (François Chollet and contributors, 2015) with TensorFlow (Abadi et al., 2016) backend. Each 3D CNN model was trained separately to predict age from a T1 MRI. The input for each network was a 3D volume, of size [90, 126, 110] and the output was a single scalar representing chronological age (years). The model was composed of two blocks, each with a batch normalization layer (Ioffe & Szegedy, 2015) followed by two 3D convolutional layers and a max-pooling layer. The two blocks were followed by two fully connected layers (FC). All layers, but the last fully connected one were followed by a ReLU nonlinear activation (Nair & Hinton, 2010). To reduce overfitting, we added dropout layers after the convolutional layer and before the last layer for the training stage (see Figure 1a for the complete architecture). The loss function for each CNN was the mean squared error between the real and predicted age. The network architecture was designed using random

TABLE 1 List of all studies which comprise the dataset

Study/database	*N*	Age *M* (±*SD*)	Gender (F;M)
Consortium for Reliability and Reproducibility (CoRR; Zuo et al., 2014)	1,378	26.0 (±15.8)	693; 685
Alzheimer's Disease Neuroimaging Initiative (ADNI; Jack et al., 2008)	1,476	73.0 (±7.0)	563; 912
Brain Genomics Superstruct Project (GSP; Buckner, Roffman, & Smoller, 2014)[a]	1,099	21.5 (±2.9)	630; 469
Functional Connectomes Project (FCP; Biswal et al., 2010)	1,067	28.9 (±13.9)	594; 473
Autism Brain Imaging Data Exchange (ABIDE; Di Martino et al., 2014)	1,053	17.1 (±8.1)	153; 900
Parkinson's Progression Markers Initiative (PPMI; Marek et al., 2011)	702	61.7 (±10.2)	260; 442
International Consortium for Brain Mapping (ICBM; Mazziotta, Toga, Evans, Fox, & Lancaster, 1995)	641	30.6 (±12.2)	293; 348
Australian Imaging, Biomarkers and Lifestyle (AIBL; Ellis et al., 2009)	616	72.9 (±6.6)	342; 273
Southwest University Longitudinal Imaging Multimodal (SLIM; Liu, Wei, Chen, Yang, & Meng, 2017)	574	20.1 (±1.3)	320; 252
Information extraction from Images (IXI; Heckemann et al., 2003)	563	48.2 (±16.5)	312; 252
Open Access Series of Imaging Studies (OASIS; Marcus, Fotenos, Csernansky, Morris, & Buckner, 2010; Marcus et al., 2007)	402	51.6 (±24.9)	257; 145
Consortium for Neuropsychiatric Phenomics (CNP; Poldrack et al., 2016)	252	33.3 (±9.3)	112; 153
Center for Biomedical Research Excellence (COBRE; Mayer et al., 2013)	146	37.0 (±12.8)	37; 109
Child and Adolescent NeuroDevelopment Initiative (CANDI; Frazier et al., 2008)	103	10.8 (±3.1)	46; 57
Brainomics (Pinel et al., 2012)	89	24.7 (±6.8)	47; 42
Overall	10,174	39.4 (±23.8)	4,659; 5,511

Note: For each study, the number of available subjects (*N*), the mean and *SD* of the age and gender distribution are provided.
[a]To prevent participant identification in the GSP study age was rounded to the closest 2 years bin.

hyperparameters search and was chosen as the one that maximized the accuracy within the validation set.

2.5 | The ensemble model

The ensemble in the current work included multiple 3D CNNs (m = 10) each trained separately to predict age from a T1 MRI. As in previous work utilizing CNNs (Lakshminarayanan, Pritzel, & Blundell, 2017; Lee et al., 2015), ensemble models differ only in their random weight initialization. Hence, they had identical architecture and were trained on the same samples. We picked m = 10 due to consideration of training time given the large number of parameters in a 3D CNN and the large training set. After each network was independently trained, a linear regression model for age prediction is learned from the outputs of the 10 networks using the same training set (see Figure 1b). The similarity of prediction error among models was tested using the signed error correlation between each two networks within the test set.

2.6 | Performance metrics

All databases were randomly divided into training (90%), validation (5%), and test (5%) sets. The training set was used to train each network separately and to find the optimal ensemble weights. The validation set was used for hyperparameters tuning, for example, augmentation and network's architecture parameters, and to assess over-fitting. All performance measures were calculated on the untouched test set. The prediction was evaluated using mean absolute error (MAE) and the Pearson correlation coefficient between the network prediction and the chronological age values.

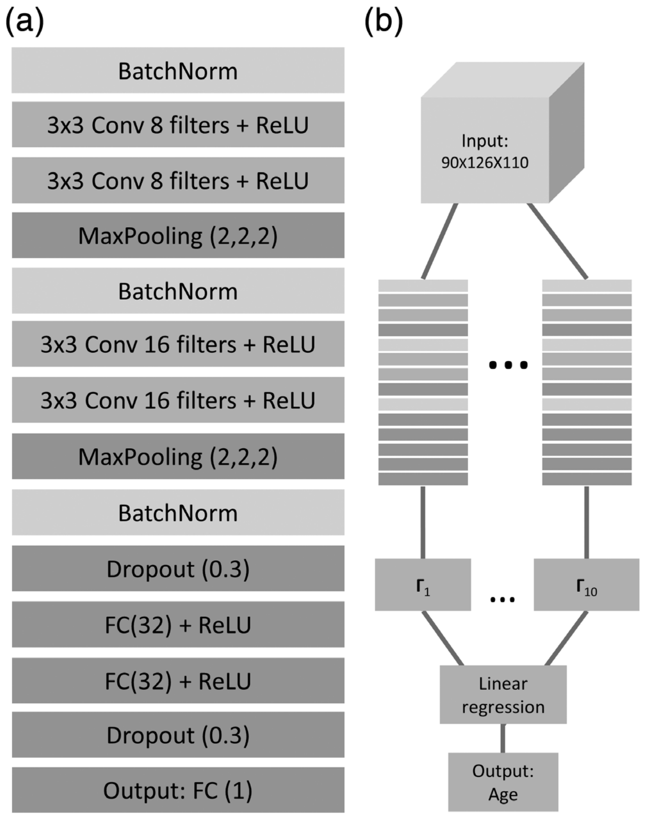

FIGURE 1 Network architecture for age prediction. (a) The detailed architecture of the network used for age prediction from 3D T1 MRI volume. BatchNorm = batch normalization, Conv = convolutional layer, ReLU = rectified linear unit, FC = fully connected layer. (b) The ensemble procedure combining the output of 10 separately trained CNNs (Γ_{1-10}) using linear regression to create the final age prediction

2.7 | Ensemble variability and uncertainty estimation

A recent work by Lakshminarayanan et al. (2017) pointed out that prediction variability within an ensemble could be utilized to evaluate uncertainty in neural networks. Here for each subject, uncertainty was quantified as the *SD* of the signed prediction error within the ensemble. In all analysis, uncertainty was evaluated within the test set. To compare uncertainty to available training sample and prediction error in different age ranges, we divided the 5–85 age range to 16 bins of 5 years each. Then, for each age range, we calculated the training sample size, mean uncertainty within the test set and the mean absolute error.

2.8 | Individual explanation maps

We employed the SmoothGrad method (Smilkov et al., 2017) that was implemented using iNNvestigate (Alber et al., 2018). This is a gradient-based method in which a given input image is first distorted with random noise from a normal distribution $N(\mu = 0, \sigma = 0.1)$, then the partial derivative of each voxel is computed with respect to the trained model's output. This was repeated several times ($k = 32$), then the produced gradient maps were averaged. We used partial derivative following Adebayo et al. (2018) work that demonstrated that it best captures the CNN's training process.

2.9 | Aggregating explanation maps across samples

First, the models preprocessed input was transformed into the raw anatomical space using FSL FLIRT (Jenkinson, Bannister, Brady, & Smith, 2002) followed by surface-based nonlinear registration to the MNI space using Freesurfer (Greve & Fischl, 2009). The transformations were computed on the T1 images, then applied to the explanation maps. The complete pipeline was created using Nipype (Gorgolewski et al., 2011). Next, each volume was standardized ($\mu = 0$, $\sigma = 1$), and smoothed with a 3D Gaussian using Scikit-image (Full width at half maximum = 4; van der Walt et al., 2014). Finally, all volumes were averaged to create population-based explanation maps. In line with previous work, we used the absolute value of the resulting maps (Ancona, Ceolini, Öztireli, & Gross, 2017). To identify regions with the highest contribution to the model's prediction, we threshold the map, keeping only 1% of the voxels with the highest gradient value (see Figure 2—for the inference scheme). To create an ensemble population-based map, we aggregate these population-based maps generated for each of the 10 CNN by taking the median of each voxel across the 10 maps. We will refer to the statistics obtained for these explanation maps, which is the standardized partial derivative, as an *explanation score* (ES).

2.10 | Assessing the similarity of explanation maps within the ensemble

To assess the diversity among independently trained CNNs, or the extent to which different CNNs utilize different brain regions for the prediction, we examined the similarity among their explanation maps. Specifically, the similarity between each pair of population-based explanation maps ($n = 100$) was evaluated with two measures: Dice similarity (Zou et al., 2004) and the modified Hausdorff distance (MHD; Dubuisson & Jain, 2002) on the threshold maps. Maps thresholding was generated by taking the absolute value of each population-based map, computing the fifth percentile of the ES within the brain mask and creating a binarized map for super-threshold values: $f(x) = \begin{cases} 1\, if\, |x| > \text{threshold} \\ 0\,\text{otherwise} \end{cases}$. For each pair of binarized maps, the Dice coefficient was defined as Dice $= \frac{2\,|X \cap Y|}{|X| + |Y|}$, where $|X \cap Y|$ is the number of overlapping super-threshold voxels in both maps, and $|X|$ and $|Y|$ are the number of super-threshold voxels for maps X and Y, respectively. MHD was derived by first finding the surface for each cluster of super-threshold voxels within each map by using a gradient-based

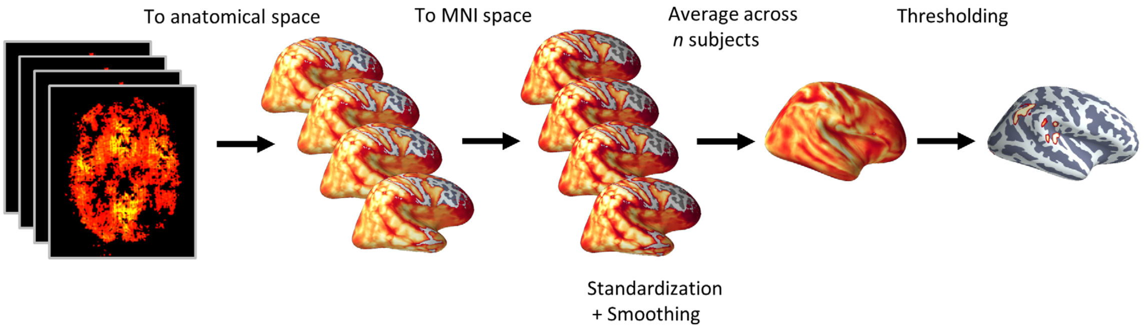

FIGURE 2 A layout of the inference scheme. For a subset of *n* subjects, an explanation map was computed, representing the contribution of each voxel to the model's output. Each saliency map was first registered to the subject anatomical image, then it was transformed to the MNI space. Next, each volume was smoothed with a 3D Gaussian. Finally, all the volumes were averaged to create a population-based explanation map

edge detector. Then, the MHD, or the symmetric average surface distance was calculated as follows:

$$\mathrm{ASD}(X,Y)=\frac{1}{\mathrm{M}_x}\sum_{k=1}^{\mathrm{M}_x}\mathrm{min}_i d(x_k,y_i)$$

$$\mathrm{MHD}(X,Y)=\frac{1}{2}(\mathrm{ASD}(X,Y)+\mathrm{ASD}(Y,X))$$

where d is the Euclidian distance, M_x is the number of voxels in the surface map X and x and y are points on the surface in maps X and Y, respectively. Both the MHD and Dice coefficients were calculated for each pair of maps creating a distance/similarity matrix. The mean distance/similarity was calculated by taking the mean over the lower triangle of that matrix.

2.11 | Relating contribution to specific tissues and brain structures

To obtain a general view of the features utilized by the model, we first segmented the brain volume to four classes of tissue type: cerebrospinal fluid (CSF) and choroid plexus, white matter (WM), subcortical gray matter (GM), and cortical GM. These were determined by applying the Desikan–Killiany Atlas (Desikan et al., 2006) using Freesurfer on the MNI template. Taking the calculated ensemble population-based explanation map, we devised a volume-normalized class ES by dividing the mean ES in each class by the total class volume. The resulting class scores were reported as a percentage of the sum of class scores. Next, to identify the specific brain regions that contributed the most to age prediction, we identified clusters of voxels in the threshold map (first percentile) using FSL cluster (Woolrich et al., 2009). For each cluster, we report the name of the brain region, its MNI coordinates, the cluster size and peak ES within the cluster. Brain regions were identified by locating the pick value of each cluster in the Desikan–Killiany Atlas for GM structures and with the ICBM-DTI-81 Atlas (Mori, Wakana, Van Zijl, & Nagae-Poetscher, 2005) for WM structures. Since many of the clusters were located within CSF spaces, whose subparts are poorly delineated in most parcellation, we manually identified subdivisions of the cisterns and ventricles. The unthresholded population-based maps for each of the 10 CNN and the ensemble map are available at Neurovault (Gorgolewski et al., 2015; https://neurovault.org/collections/5552/).

2.12 | Validating the population-based inference scheme

2.12.1 | Replicability of the produced explanation map as a function of sample size

To examine whether creating explanation maps based on a larger population would increase the split-sample similarity, two population-level explanation maps were created by sampling *m* subjects (*m* = 1, 6, 11, …, 101) with replacement from two groups. The groups were created by randomly splitting to half, a sample of 200 subjects from the test and validation sets. Each map was thresholded, binarized (see Section 2.8) and the Dice similarity and the MHD were calculated between the two maps as a function of the sample size *m*. The procedure was repeated 100 times, and for each iteration, the 200 subjects were randomly assigned to the two groups. This test was repeated independently for each population-based explanation map derived from the 10 CNNs.

2.12.2 | Similarity between explanation maps and voxel-based morphometric meta-analysis

To examine whether the derived explanation map elicits similar regions to those detected with established methods, we compared it with a baseline obtained from studies that use voxel-based morphometry (VBM; Ashburner & Friston, 2000) to test structural age-related changes. Briefly, in the VBM method, a mass-univariate test between the tissue composition of any voxel in the brain and a given external variable (age) is conducted. To address the differences in brain position and anatomy, all brain volumes were normalized to a common space and then smoothed by a Gaussian kernel to account for small registration differences. In the current study, we used a published activation likelihood estimation (ALE) meta-analysis of age VBM studies (Vanasse et al., 2018). Here, by utilizing peak reported coordinates from several VBM studies (*n* = 43), the ALE analysis assigns each voxel the probability that it lies within a reported peak (Laird, Bzdok, Kurth, Fox, & Eickhoff, 2011). The ALE value across all the superthreshold (first percentile) voxels in the ensemble population-based explanation map was averaged. This empirical value was compared to a null distribution created by randomly sampling 1% of the voxels within the brain mask.

2.12.3 | Specificity of the regions obtained in the analysis to the employment of the current model

To evaluate the contribution of the regions discovered using the population-based explanation map to the prediction of the current model, we examined how variability in their age-controlled volume correlated with the model's prediction error. Regional volumes of cortical and subcortical areas were extracted using the Desikan-Killiany atlas (Desikan et al., 2006) computed using Freesurfer following by regressing out the total intracranial volume (Voevodskaya et al., 2014). Next, subjects' chronological age was further regressed out from these values to produce the age-normalized volume. Prediction error was formulated as the signed difference between the chronological age and the predicted age. The test was conducted separately for each anatomical ROI in the parcellation. Specific ROIs within the Desikan–Killiany parcellation, such as WM hyperintensities and the fifth ventricle, exist for only some of the subjects (*n* = 619; from the test/validation sets), and thus were subsequently excluded (9 excluded, 98 remained; see Figure S6 for the complete list).

3 | RESULTS

We start by presenting the model's ensemble performance for predicting subject chronological age from their T1 structural images on an unseen test set (*N* = 526). Then, using a novel inference scheme, we locate the anatomical regions that contributed the most to the model's prediction. We validate the robustness of our inference scheme in three ways. First, we demonstrate that it substantially increases reliability compared to previous methods, creating more coherent and localized explanation maps. Second, we quantitatively compare these explanation maps to age voxel-based morphometric studies, demonstrating significant overlap with a simple baseline model. Finally, we demonstrate that this approach enables to gain specific insights about the model by identifying brain regions for which the model exhibits the highest correlation to inter-subject volumetric variability.

3.1 | Estimating "brain age"

Several attempts were previously made to identify the relation between chronological age and brain structure, using various feature extraction techniques, advanced preprocessing methods and a relatively limited sample size (Irimia, Torgerson, Goh, & Van Horn, 2015; Kandel, Wolk, Gee, & Avants, 2013; Shamir & Long, 2016). Here, we build upon recent progress in utilizing CNNs for predicting chronological age from raw structural brain imaging (Cole & Franke, 2017) and introduce substantial improvements using an ensemble of models. In the current work, 10 randomly initialized CNNs were separately trained. The mean MAE across networks was 3.72 years (±0.17), and the Pearson correlation between the predicted and the chronological age was 0.97 (±0.001). Next, a simple linear regression model was trained on the output of each network to find an optimal linear combination between them, yielding an MAE of 3.07 and a Pearson correlation of 0.98 to the chronological age (Figure 3; see Figure S1 for evaluation per dataset). To demonstrate the prediction similarity among all CNNs, evident in the ensemble prediction gain, we tested the signed error correlation between each two networks within the test set. The mean correlation coefficient value was 0.73 and the *SD* was 0.09 (see Figure S8). A model dataset-specific bias was found only for the SLIM dataset, in which age was over-estimated in 1 year

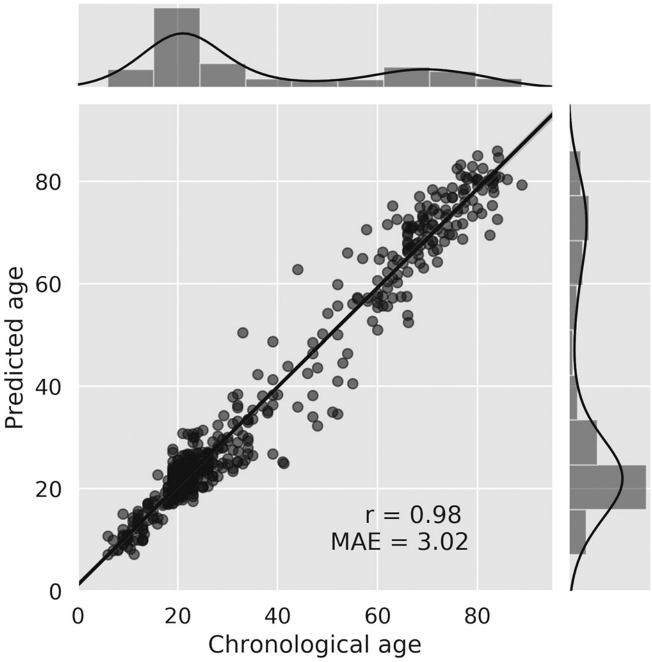

FIGURE 3 Regression plot of the chronological age compared to the model's prediction for the test set. The main plot depicts the Pearson correlation coefficient between the chronological and the predicted age; the Pearson correlation coefficient (*r*) and the mean absolute error (*MSE*) are indicated on the plot. The data points are presented with partial transparency thus overlapping points are shown in darker gray. The top and right panels of the figure depict histograms and kernel density plots of the distribution of the chronological age and the predicted age (respectively) obtained in the test set

(t = 2.61, p = .01). All other datasets did not present such bias (all t's < 1.51, all p's > .14; see Figure S9). An analysis of MAE as a function a training sample and age range is included in the supplementary section (Figures S10 and S11, respectively).

3.2 | Ensemble variability and uncertainty estimation

Apart from a prediction gain, the use of an ensemble provides a simple way to evaluate prediction uncertainty (Lakshminarayanan et al., 2017). Model uncertainty can be viewed as the lack of confidence in the prediction made, given the inability of the model to capture the true data generating process. Here, uncertainty was measured by calculating the prediction variability within the ensemble. Since the uncertainty metric aims to evaluate the lack of confidence in the prediction, it is expected to be correlated with the prediction error (Kendall & Gal, 2017). Accordingly, we found a significant correlation between the MAE and uncertainty (r = 0.398, p < .001; Figure S12). We additionally examined whether uncertainty would be higher for age ranges where less training data is available (Gal & Ghahramani, 2016). We found the maximum uncertainty in the 30–35 age range where the availability of data samples was more limited. The opposite was found for the 10–25 and the 65–75 age ranges that had the largest sample size (see Figure S11).

3.3 | From the model to the brain—A novel inference scheme

Building upon previous attempts to assign pixel-wise (or voxel-wise) explanation measures to a model's prediction (Smilkov et al., 2017), we propose that creating explanation maps based on a population, rather than on a specific sample, may substantially improve the coherence and reliability of these maps. We create these maps for each of the 10 independently trained CNNs and examine their similarity. Then, using an aggregated map across all 10 networks we present the brain regions that contributed the most to predicting age.

3.3.1 | Assessing the similarity of explanation maps within the ensemble

Explanation maps for the 10 CNNs, averaged across 100 subjects from the test/validation set were produced to create a population-based map (see Section 2.7; Figure S2). First, to assess the similarity between each pair of these produced maps among the 10 independently trained networks, each map was thresholded (fifth percentile) and binarized, then the Dice coefficient similarity measure was computed for each pair of maps (Figure S3). We found a significant Dice similarity across all 45 possible pairs (Dice coefficient: m = 0.17, SD = 0.058; binomial test: p < .001). Since Dice similarity fails to capture the relation between two maps that are adjacent in the Euclidean space but nonoverlapping, we additionally computed the MHD (Dubuisson & Jain, 2002), taking the symmetric average surface distance, among all pairs. We found that the mean MHD among all possible pairs was 6.44 mm (SD = 1.22). Thus, even though these different population-based maps were derived from independently trained networks, there is a moderate, significant, overlap between them. The fact that this overlap is merely moderate coincides with the prediction differences that allow the accuracy gain in ensemble prediction.

3.3.2 | Mapping the anatomical regions underlying "brain age" prediction

After estimating the similarity among the different explanation maps for the 10 CNNs, we created an ensemble population-based map by taking the median value for each voxel across all networks. We report how the ES is distributed among different tissue types, and among different anatomical regions in order to examine their contribution for age prediction. Testing the volume-normalized contribution of each tissue type, we found that cavities containing CSF and choroid plexus had the highest contribution (35.62%), followed by subcortical GM (27.66%), WM (19.49%), and finally, cortical GM (17.23%) which contributed the least. Table 2 presents the location of clusters (>100 voxels) in the threshold explanation map (first percentile). We found that the structures contributing most to age prediction in our model were the ventricles, subarachnoid cisterns, and their borders (see Figure 4). Specifically, the fourth ventricle, the ambient cistern bilateral to the midbrain, the superior cerebellar cistern, the bilateral Sylvian cistern, the lateral ventricles, the interpeduncular cistern, and the right parahippocampal fissure. WM tracks that were found important for age prediction were the bilateral tapetum, the right anterior limb of the internal capsule and the left medial lemniscus. Finally, the bilateral thalamus and the right precentral gyrus were the GM regions that contributed most to the prediction. Both of these analyses support the notion that age prediction in the current model is largely based on age-related morphological changes in the cavities containing CSF.

3.4 | Validating the population-based inference scheme

To validate the suggested approach for detecting regional contribution to a CNNs ensemble, we conducted three tests. First, we tested the importance of sample size to the explanation maps replicability by testing the half-split similarity of these maps as a function of the population size. Second, to test whether these results coincide with data from other studies, we tested the extent of the similarity between the produced maps and a meta-analysis of VBM studies. Lastly, to confirm the specificity of these results to the current model, we examined whether the produced maps highlight the particular brain regions for which the model exhibits the highest correlation to inter-subject volumetric variability.

TABLE 2 Anatomical location of clusters in the threshold explanation map

		MNI coordinates			
Region	Cluster size	x	y	z	Peak ES
4th ventricle and ambient cistern	5,531	-2	−43	−39	4.71
Superior cerebellar cistern	4,619	−3	−55	0	2.54
R tapetum	1,984	27	−43	18	1.35
L Sylvian cistern	1,409	−45	−16	11	1.62
R lateral ventricle	1,115	6	1	8	1.42
R Sylvian cistern	1,105	41	−20	0	1.56
R lateral ventricle	1,024	30	−49	2	1.83
R anterior limb of internal capsule	847	11	6	2	1.48
3rd ventricle	787	0	−26	11	1.89
L lateral ventricle	740	−28	−52	4	2.08
Interpeduncular cistern	446	1	−17	−22	2.09
R Sylvian cistern	440	39	12	−20	1.23
L medial lemniscus	307	−2	−36	−41	1.25
L thalamus	303	−12	−18	11	0.993
L ambient cistern	238	−12	−34	−13	1.18
Tapatum left	231	−26	−49	15	0.989
R thalamus	212	13	−17	−2	0.925
L ambient cistern	171	−16	−24	−21	1.17
R precentral gyrus	127	50	11	29	0.958
Perihippocampal fissure	110	32	−9	−26	1.06

Note: MNI coordinates of clusters in the ensemble population-based explanation map threshold for the first percentile. Cluster size is reported in terms of the number of voxels, and the peak ES is the maximum ES within the cluster.

3.4.1 | Replicability of the produced explanation map as a function of sample size

It is not clear to what extent an explanation map derived from a single sample would indeed represent the entire population. To examine this issue, we tested the split-sample similarity of the explanation maps with a gradually increasing sample size obtained from two separate groups. In each test repetition (k = 100), the groups were created by randomly half-splitting a sample of 200 subjects (see Section 2.10). We report the Dice similarity and the MHD among these maps as a function of the sample size drawn from them (see Figure 3). Across all 10 networks, we found an increase of the Dice similarity and a decrease of the MHD as a function of the sample size, ranging from a single sample to 101 samples (mean Dice = 0.19, mean MHD = 3.73; mean Dice = 0.74, mean MHD = 1.00; respectively) (Figure 5a,b). The relative improvement in the replicability of these maps asymptotes at 40–60 subjects, such that adding more samples had little further impact. Figure 5c shows 2D glass brain projections of the population-based maps to illustrate the change as a function of the sample size, resulting in a visually apparent increase in coherence and a decrease in noise. These results suggest that whether due to noise or fundamental differences in subject-specific maps, an explanation produced from a single sample rather than a population has low replicability.

3.4.2 | The similarity between explanation maps and voxel-based morphometric meta-analysis

To quantitatively assess whether the regions detected in the current inference scheme coincide with previous findings, we compared the resulting explanation maps to data obtained from VBM studies testing structural age-related changes. The VBM method has the desired property of allowing to test the relation between the estimated tissue composition of each voxel and any given relevant variable, age in the present case. In contrast, other methods are limited to a specific set of ROIs or a given brain parcellation that often fails to properly parcellate non-GM regions. We used a published activation likelihood estimation (ALE) meta-analysis of age VBM studies (n = 15, Vanasse et al., 2018) in which each voxel is assigned with a probability for its location in a reported peak coordinate in one of the studies (Laird et al., 2011). Using this map, we examined whether that ALE value is significantly higher within the regions identified using the threshold explanation map with a permutation test. The mean ALE value within the super-threshold (first percentile) explanation map was higher than any set of randomly selected voxels in the permutation test (k = 10,000; meta-analysis: empirical mean ALE: 0.003, p < .0001; see Figure S4). Interestingly, both methods highlighted regions surrounding the lateral and third ventricles, subcortical areas, and the bilateral insula/Sylvian cistern, as opposed to cortical regions that appeared only in the ALE map (see Figure S5).

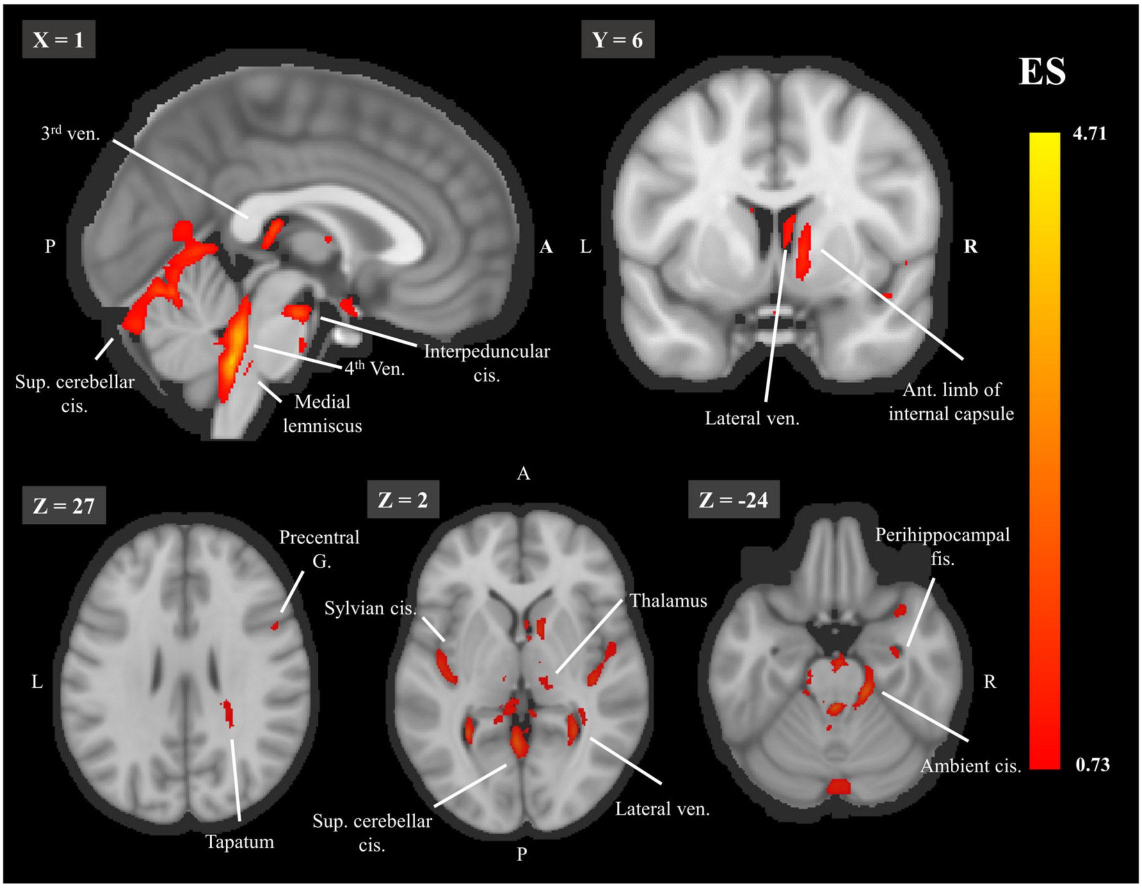

FIGURE 4 The threshold explanation map shown on a midsagittal (top left), a coronal (top row left) and 3 axial (bottom row) slices. Aggregated explanation map across 100 subjects and the 10 networks, thresholded for the first percentile of the ES. Abbreviations: ant. = anterior, cis. = cistern, g. = gyrus, fis. = fissure, ven. = ventricle. For each image, the slice number in the MNI template is indicated on the left upper corner. The color bar indicates the values of the ES

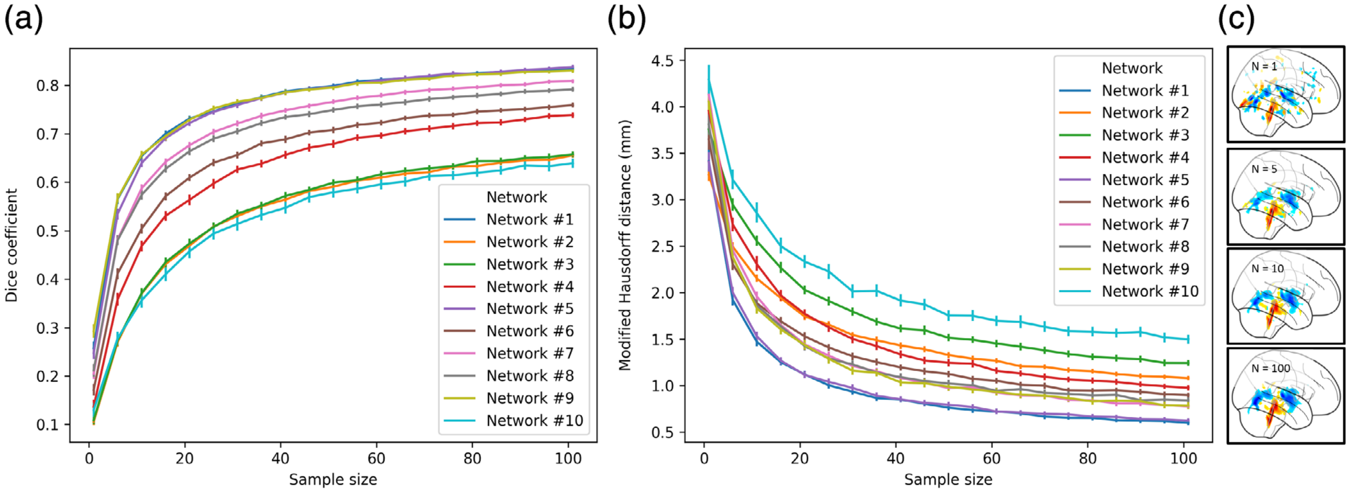

FIGURE 5 The split-sample similarity of the explanation maps as a function of sample size. The similarity of two maps produced from an increased sample size from two separate groups (n = 100 for each group) was measured using (a) Dice coefficient and (b) MHD (mm). The results are reported for all 10 CNNs and each is presented in a separate color. (c) A visual illustration of an explanation map for network 1 produced by increasing the sample size (from top to bottom, N = 1,510,100). The error bars represent a 95% confidence interval

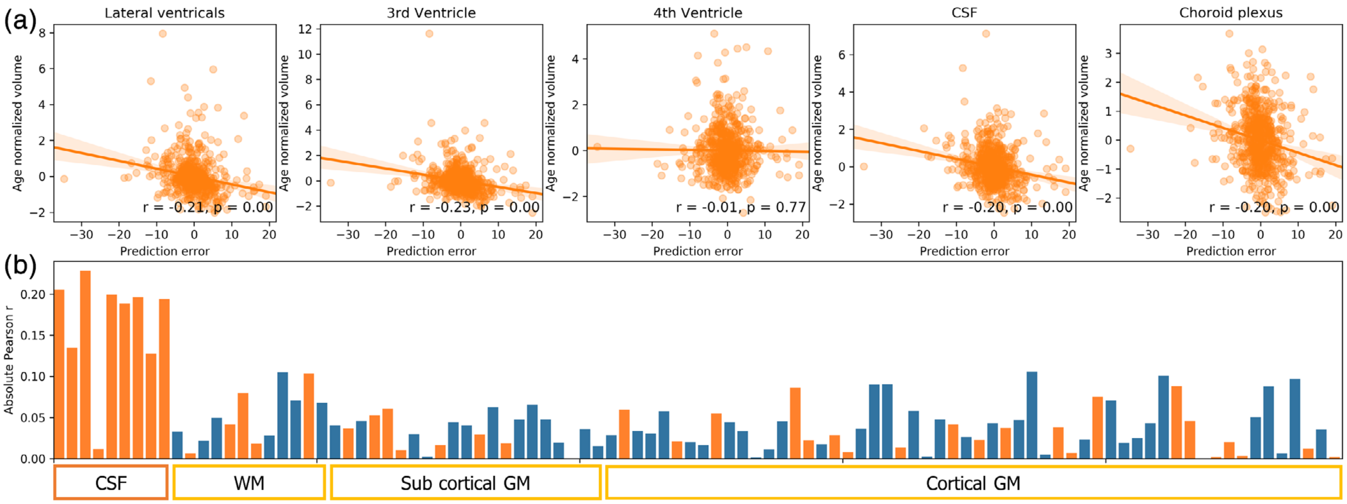

FIGURE 6 Deviation in volume from age norm and prediction error. (a) Graphs of five ROIs, detected with the current inference scheme, showing the correlation between the age-controlled volume and the signed prediction error. Age-normalized volume was computed by regressing out subjects chronological age from the measured volume. Volume was determined according to the Desikan-Killiany atlas fitted with Freesurfer. Prediction error was formulated as the chronological age minus the predicted age. Note that for the sake of brevity, in the upper five plots, the volume of the lateral ventricles and choroid plexus was computed as the sum of their subparcellations. (b) The bar graph depicts the correlation between the age normalized volume and the signed prediction error for all the 98 regions in the parcellation. Positive correlations are presented in blue and negative in orange for simple magnitude comparison. As shown, the age-controlled volume of cavities containing CSF and the choroid plexus (L/R Lateral Ventricle, L/R inferior Lateral Ventricle, 3rd ventricle, nonventricles CSF, L/R choroid plexus), except for the 4th ventricle, had the largest correlation with the model's prediction error compared with all other WM/GM regions (see Figure S6 for the full labels)

3.4.3 | Specificity of the regions obtained in the analysis to the employment of the current model

Prediction errors could result from the inability of the model to capture the complexity of the brain aging process or due to the natural variability in brain morphology within the population. Exploiting the latter, in the current analysis, we aimed to examine whether prediction error would correlate with volumetric variability of specific brain regions. Specifically, by applying the Desikan–Killiany atlas using Freesurfer, we tested whether age-controlled volume of the ventricles and cisterns that were highlighted by the inference scheme were correlated with the CNNs ensemble prediction error. Indeed, we found a significant correlation between the age-normalized volume and the prediction error for the ventricles excluding the fourth ventricle, the choroid plexus and nonventricular CSF (*n* = 619; for all 9 regions but fourth ventricle: $r > 0.13$, $p < .002$). This correlation was higher in these regions than in any other brain region in the parcellation supporting the specificity of the results to the regions obtained using the population-based explanation maps (see Figure 6). Interestingly, this specificity was not apparent when examining the correlation between regional volume and chronological age, in which significant correlation is seen in almost all regions (>93% of the ROI; see Figure S7). Put differently, while the volume of almost all regions correlated with age, deviance from the age norm in the ventricles and CSF, detected using the population-based maps, best reflected the prediction error. This suggests that the current inference scheme not only detected regions that are altered in aging, but it also detected the distinct regions that had the highest contribution to the current prediction, attesting to the high specificity of this method to the applied model.

4 | DISCUSSION

In the current study, we examined whether individuals' chronological age could be predicted from T1 MRI scan and whether it is possible to localize the underlying brain regions that allow such prediction. Using a large aggregated sample of 10,176 subjects we trained and validated an ensemble of 3D CNN models, and showed that "brain age" could be estimated from raw T1 MRI with MAE of ~3.1 years. We demonstrated that the use of an ensemble of models rather than a single estimator reduces the MAE in more than 6 months and provided evidence that such gain is due to the difference in the features or brain regions that are utilized by each model. Models ensemble additionally allowed a simple estimation of model uncertainty. We found that model uncertainty correlated with prediction error and was higher in age ranges where less training data was available. Brain age was previously shown to be indicative of neurodegenerative diseases and other clinical conditions (Cole & Franke, 2017), thus improving the precision, confidence estimation, as well as the interpretability of this biomarker, could be an important step toward integrating it in clinical use.

4.1 | Identifying the brain regions underlying age prediction using population-based explanation maps

Drawing from previous studies aiming at identifying regional contributions to the model's prediction, we aimed to locate the brain regions that governed our brain age estimation. Here we presented a novel approach to aggregate multiple explanation maps from several subjects, thus creating a population-based map. This was achieved by deriving a series of transformations warping the 3D volumes presented to the CNN into the MNI space. We then applied those transformations to the computed explanation maps, thus allowing to average different explanation maps in a common space. This approach precludes the need for pre-registration to a common template in the training stage, as done previously (Ceschin et al., 2018), a step that is error-prone, time-consuming and might result in the loss of relevant structural information (Iscan et al., 2015). Thus, these maps are obtained without compromising predictive accuracy, since the model's training objective is not altered.

To validate our method, we tested how it affects three important aspects. First, we quantitatively assessed how sample size in population-based maps improved their reproducibility. We reported a substantial improvement in split-sample similarity as moving from a map based on a single subject to a map based on a population of 40–60 subjects. The low split-sample similarity of single-subject maps emphasized the need to apply such practices when analyzing these explanation maps. Next, we demonstrated that despite the methodological differences, the proposed map exhibited significant similarity to ALE maps obtained from an age VBM meta-analysis study (Vanasse et al., 2018), attesting to its convergence validity. Finally, using regional volumetric measures, we demonstrated that the brain regions highlighted by our method were those with the highest correlation to the model's prediction error, indicating the specificity of the derived maps to the current model.

4.2 | Reducing noise or averaging over true relevant population differences?

Comparing our approach for deriving population-based explanation maps to an approach based on a single sample as in Smilkov et al.'s (2017) paper, we demonstrate an increase in reproducibility and a distinct visual improvement in the coherence of these maps. We therefore discuss possible mechanisms that may account for these findings. In their study, Smilkov et al. (2017) demonstrated that the derivative of CNNs are highly noisy, and averaging explanation maps obtained from several noised samples of the same input can improve these maps. Here, after applying the Smilkov et al. (2017) method, we further averaged multiple explanation maps derived from different inputs, that is, brain volumes of different subjects. A possible explanation to the apparent reproducibility improvement is that sampling from the true input distribution (brain volumes of different individuals), rather than mere noised samples of the same input, would result in estimation that is more robust to local gradient noise. A second, nonexclusive possible account might be that the model was trained on brain volumes from a heterogenic population. Differences in brain aging trajectory were found both at the individual level (Raz, Ghisletta, Rodrigue, Kennedy, & Lindenberger, 2010) and among different populations, for example, in relation to gender (Jäncke et al., 2015). Thus, it is possible that the model extracts different features due to relevant structural variability in different populations.

4.3 | The ventricles and cisterns as biomarkers for brain aging

Aging is accompanied by multiple processes affecting the human brain, manifested in structural changes that could in part be quantified by neuroimaging (see Lorio et al., 2016). Accordingly, a wealth of literature reported a complex pattern of morphological changes evident across all brain regions, but arguably more apparent in some areas such as the frontal lobes, insular cortices, and the hippocampus (Fjell et al., 2009). Previous work that examined the predictive value of voxel-wise feature maps reported that features derived from GM, compared to nonbrain tissue, served as better age predictors (Monté-Rubio, Falcón, Pomarol-Clotet, & Ashburner, 2018). Interestingly, in our model, the ventricles and cisterns were highlighted as most relevant for age prediction. Several possible reasons might account for this finding. First, CSF volume was found to increase already from young adulthood (Courchesne et al., 2000), thus it may constitute an early aging biomarker. Since CSF pressure remains relatively constant and even decreases in old age (Fleischman et al., 2012), it is likely that CSF expansion reflects a decrease in WM/GM volumes rather than an increase in CSF pressure. Thus, CSF volume changes might be a surrogate for general brain atrophy, as suggested in previous work (De Vis et al., 2016). In line with this account, CSF volume was previously found as a better aging marker compared to WM/GM/hippocampal volume (Vinke et al., 2018). Notably, CNN representations are learned from the raw data and can potentially identify morphological alterations in these regions that facilitate aging prediction. In contrast, previous models (Liem et al., 2016; Valizadeh, Hänggi, Mérillat, & Jäncke, 2017) exclusively based on regional volumetric measurements presented substantially lower accuracy. It is important to stress that the created explanation maps do not directly highlight regions that are indicative of age. Instead, given a specific model and a set of images, the maps highlight regions that are likely to contribute to the model's prediction. Accordingly, it has been previously suggested that neural networks present an inductive bias to more simple or parsimonious solutions (Neyshabur, Tomioka, & Srebro, 2015). Thus, it is possible that although brain regions other than the ventricular system are indicative of age, the saliency of the ventricles and cisterns, due to their high contrast, allows their capture by the network. These possible reasons could be tested in future work but nevertheless, the ability to generate new biologically relevant hypotheses from a deep learning predictive model is a desirable practice supported here by our novel inference scheme.

4.4 | Ensemble diversity among models' population-based explanation maps

Evidence suggests that prediction based on a set of learning algorithms instead of a single algorithm will result in an accuracy gain (Sagi & Rokach, 2018) that increases as these models are more accurate and diverse (Breiman, 2001; Kuncheva & Whitaker, 2003). Learning diverse models could be achieved by changes in architecture (Singh, Hoiem, & Forsyth, 2016) or introducing different subsets of the training data to each model (Benou, Veksler, Friedman, & Riklin Raviv, 2017). In the context of deep CNN, as opposed to convex or shallow learning algorithms, it has been shown that models that differ only in their random weight initializations, constitute an ensemble that is not only adequately diverse, but performs better than models exposed to different subsets of the data (Lakshminarayanan et al., 2017; Lee et al., 2015). In the current work, we examined the similarity among pairs of population-based explanation maps derived from different models within such an ensemble. Although within each model population maps showed high reliability, on average, pairs of models exhibit only moderate similarity. This supports the notion that random weight initializations generate diverse models that utilized different parts of the input, that is, different brain regions. This might explain the observed improvement in prediction accuracy when using an ensemble. The apparent variability of explanation maps within the ensemble could be additionally considered in terms of uncertainty. Ensembles were previously utilized to evaluate uncertainty for regression and classification problems (Lakshminarayanan et al., 2017), for evaluating actions in reinforcement learning (Gal & Ghahramani, 2016) and in estimating voxel-wise uncertainty in diffusion imaging super-resolution (Tanno et al., 2019). In line with this, we suggest that the variability of voxel-wise explanation maps could be similarly viewed as confidence in the importance of various regions for a given model architecture and a training set. Thus, regions such as the ambient and cerebellar cisterns, consistently utilized across all models, could be viewed as important for the prediction with higher confidence. Overall, it seems that general conclusions regarding the contribution of different brain regions to age prediction should be made based on maps converging from multiple models.

4.5 | Limitations

This study has a number of limitations. First, it is unclear whether the trained model is fully invariant to variables such as scanning site or acquisition parameters. When testing for bias in the dataset level, each dataset with its own scanning parameters, we found evidence for a systematic prediction bias in only one of 11 studies. It is still possible, however, that the network could distinguish between scans based on their acquisition statistics and utilize such information for the prediction. This issue should be further examined in future work (see Tzeng, Hoffman, Saenko, & Darrell, 2017 for domain invariance in machine learning). Second, since the model was trained solely on cross-sectional data, it only gained information on between-subjects aging variability. Incorporating longitudinal data can allow to model individuals' aging trajectories. Finally, CNNs are complex functions with hundreds of thousands of parameters and multiple layers with nonlinearities, thus they could not be fully reduced to a set of local contributions. CNNs are likely to entail complex multivariate interactions that are not necessarily local. Hence it is important to state that our maps, based on partial derivatives, are merely an approximation of the significance of various input regions.

4.6 | Population-level explanation maps: Future directions

Computing population-based explanation maps allow examining group differences in maps produced from different populations. For example, one might ask whether a CNN model would extract different aging biomarkers for men versus women or for healthy elderly versus individuals diagnosed with AD. These tests could be applied on maps derived from two identical models separately trained on different populations or within the same model trained on both populations. In the latter case, subjects' group affiliation could be explicitly introduced to the model as an input. Alternatively, it will be possible to test whether a distinction among populations in the form of explanation maps differences, would arise without introducing such an input. Hence, explanation maps obtained from a population of subjects, registered to the same template could allow harnessing known neuroscience statistical procedures based on voxel or regional wise comparison of within compared to between-group variability. Another possible extension of the current work is the adoption of population-level explanation maps to other neuroimaging prediction problems. An example of such potential usage could be a deep learning decoding model of neural activity (Beliy et al., 2019), a model predicting the presence of a neurological condition (Li, Liu, Sun, Shen, & Wang, 2018), or any machine learning application based on a differentiable function. This, of course, would require a pre-trained model for a given task, and relies on the assumption that the model exploits regional features for the prediction.

4.7 | Conclusions

Incorporating deep learning for analysis of neuroimaging data requires improvement in both the accuracy of these predictive models and the ability to interpret them, as we aimed to address in the context of age prediction. Respectively, in the current work, we demonstrated that an individual's chronological age could be estimated with a MAE of 3.1 years from their raw T1 images, yielding a robust biomarker across several datasets. We further showed that aggregating multiple explanation maps substantially increases their reproducibility and allow to create a coherent and localized map depicting and quantifying the contribution of different brain regions to age prediction. From these maps, we conclude that the ventricles and cisterns govern these predictions. We argue that this ability to pinpoint specific brain areas is a key step for utilizing these models as possible brain health biomarkers.

ACKNOWLEDGMENTS

This research was supported by an Internal funding grant for interdisciplinary research in Ben Gurion University of the Negev to GA and IS. Our project would not have been possible without several open databases and groups that invested considerable resource and efforts to support neuroimaging data sharing. We wish to acknowledge those study groups and funding agencies: *CORR*—Data were provided in part by the Consortium for Reliability and Reproducibility (http://fcon_1000.projects.nitrc.org/indi/CoRR/html/index.html). *ADNI*—Data collection and sharing for this project was 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). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. *GSP*—Data were provided in part by the Brain Genomics Superstruct Project of Harvard University and the Massachusetts General Hospital, (Principal Investigators: Randy Buckner, Joshua Roffman, and Jordan Smoller), with support from the Center for Brain Science Neuroinformatics Research Group, the Athinoula A. Martinos Center for Biomedical Imaging, and the Center for Human Genetic Research. Twenty individual investigators at Harvard and MGH generously contributed data to GSP. *FCP*—Data were provided in part by the Functional Connectomes Project (https://www.nitrc.org/projects/fcon_1000/). *ABIDE*—Primary support for the work by Adriana Di Martino, and Michael P. Milham and his team was provided by the NIMH (K23MH087770), the Leon Levy Foundation, Joseph P. Healy and the Stavros Niarchos Foundation to the Child Mind Institute, NIMH award to MPM (R03MH096321), National Institute of Mental Health (NIMH 5R21MH107045), Nathan S. Kline Institute of Psychiatric Research), Phyllis Green and Randolph Cowen to the Child Mind Institute. *PPMI*—Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI—a public-private partnership—is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including [list of the full names of all of the PPMI funding partners can be found at www.ppmi-info.org/fundingpartners]. *ICBM*—Data used in the preparation of this work were obtained from the International Consortium for Brain Mapping (ICBM) database (www.loni.usc.edu/ICBM). The ICBM project (Principal Investigator John Mazziotta, M.D., University of California, Los Angeles) is supported by the National Institute of Biomedical Imaging and BioEngineering. ICBM is the result of efforts of co-investigators from UCLA, Montreal Neurologic Institute, University of Texas at San Antonio, and the Institute of Medicine, Juelich/Heinrich Heine University—Germany. *AIBL*—Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of aging (AIBL) funded by the Commonwealth Scientific and Industrial Research Organization (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au. *SLIM*—Data were provided by the Southwest University Longitudinal Imaging Multimodal (SLIM) Brain Data Repository (http://fcon_1000.projects.nitrc.org/indi/retro/southwestuni_qiu_index.html). *IXI*—Data were provided in part by the IXI database (http://brain-development.org/). *OASIS*—OASIS is made available by Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN). Support for the acquisition of this data and for data analysis was provided by NIH grants P50 AG05681, P01 AG03991, P20 MH071616, RR14075, RR 16594, U24 RR21382, the Alzheimer's Association, the James S. McDonnell Foundation, the Mental Illness and Neuroscience Discovery Institute, and HHMI. *CNP*—Data used in the preparation of this article were obtained from the Consortium for Neuropsychiatric Phenomics (NIH Roadmap for Medical Research grants UL1-DE019580, RL1MH083268, RL1MH083269, RL1DA024853, RL1MH083270, RL1LM009833, PL1MH083271, and PL1NS062410). This data was obtained from the OpenfMRI database. Its accession number is ds000030. *COBRE*—Data were provided by the Center for Biomedical Research Excellence (COBRE) (http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html). *CANDI*—Data were provided in part by the Child and Adolescent NeuroDevelopment Initiative (CANDI; Kennedy et al., 2012)—Schizophrenia Bulletin 2008 project. *Brainomics*—Data were provided in part by the Brainomics project (http://brainomics.cea.fr/). *NITRC*—Data used in this manuscript was partially accessed through the Neuroimaging Informatics Tools and Resources Clearinghouse. NITRC is funded by the NIH Grant numbers: 2R44NS074540 and 1U24EB023398a. *IDA*—Data used in this manuscript was partially accessed through the LONI Image and Data Archive. IDA is funded by the NIH and the NIBIB grant numbers P41EB015922 U54EB020406.

DATA AVAILABILITY STATEMENT

All the datasets used for the model's training, validation and testing were acquired from open-access data sharing projects. The results of the main analysis, the unthresholded population-based maps for each of the ten CNN and the ensemble map are available at Neurovault (Gorgolewski et al., 2015; https://neurovault.org/collections/5552/).

ORCID

Gidon Levakov https://orcid.org/0000-0002-5520-3556

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Zheng, X. (2016). TensorFlow: Large-scale machine learning on heterogeneous distributed systems. https://doi.org/10.1038/nn.3331

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 9505–9515.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., … Kindermans, P.-J. (2018). iNNvestigate neural networks! ArXiv. Retrieved from http://arxiv.org/abs/1808.04260

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2017). A unified view of gradient-based attribution methods for Deep Neural Networks. In *31st Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 1–16). ETH Zurich.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—The methods. *NeuroImage*, *11*(6 I), 805–821. https://doi.org/10.1006/nimg.2000.0582

Barnes, J., Bartlett, J. W., van de Pol, L. A., Loy, C. T., Scahill, R. I., Frost, C., … Fox, N. C. (2009). A meta-analysis of hippocampal atrophy rates in Alzheimer's disease. *Neurobiology of Aging*, *30*, 1711–1723. https://doi.org/10.1016/j.neurobiolaging.2008.01.010

Beliy, R., Gaziv, G., Hoogi, A., Strappini, F., Golan, T., & Irani, M. (2019). From voxels to pixels and back: Self-supervision in natural-image reconstruction from fMRI. *Advances in Neural Information Processing Systems*, 6514–6524. Retrieved from. http://www.wisdom.weizmann.ac.il/~vision/ssfmri2im/

Benou, A., Veksler, R., Friedman, A., & Riklin Raviv, T. (2017). Ensemble of expert deep neural networks for spatio-temporal denoising of contrast-enhanced MRI sequences. *Medical Image Analysis*, *42*, 145–159. https://doi.org/10.1016/J.MEDIA.2017.07.006

Bermudez, C., Plassard, A. J., Chaganti, S., Huo, Y., Aboud, K. S., Cutting, L. E., … Landman, B. A. (2019). Anatomical context improves deep learning on the brain age estimation task. *Magnetic Resonance Imaging*, *62*, 70–77. https://doi.org/10.1016/j.mri.2019.06.018

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., … Milham, M. P. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(10), 4734–4739. https://doi.org/10.1073/pnas.0911855107

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buckner, R. L., Roffman, J. L., & Smoller, J. W. (2014). Brain genomics Superstruct project (GSP). *Harvard Dataverse*, *10*. https://doi.org/10.7910/DVN/25833

Cardenas, V. A., Chao, L. L., Studholme, C., Yaffe, K., Miller, B. L., Madison, C., … Weiner, M. W. (2011). Brain atrophy associated with baseline and longitudinal measures of cognition. *Neurobiology of Aging*, *32*(4), 572–580. https://doi.org/10.1016/j.neurobiolaging.2009.04.011

Ceschin, R., Zahner, A., Reynolds, W., Gaesser, J., Zuccoli, G., Lo, C. W., & Gopalakrishnan, V. (2018). A computational framework for the detection of subcortical brain dysmaturation in neonatal MRI using 3D convolutional neural networks. *NeuroImage*, *178*, 183–197. https://doi.org/10.1016/j.neuroimage.2018.05.049

Cole, J. H., & Franke, K. (2017). Predicting age using: Neuroimaging: Innovative brain ageing biomarkers. *Trends in Neurosciences*, *40*(12), 681–690. https://doi.org/10.1016/j.tins.2017.10.001

Cole, J. H., Poudel, R. P. K., Tsagkrasoulis, D., Caan, M. W. A., Steves, C., Spector, T. D., & Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, *163*, 115–124. https://doi.org/10.1016/j.neuroimage.2017.07.059

Cole, J. H., Ritchie, S. J., Bastin, M. E., Hernández, M. V., Maniega, S. M., Royle, N., … Wray, N. R. (2018). Brain age predicts mortality. *Molecular Psychiatry*, *23*(5), 1385–1392.

Courchesne, E., Chisum, H. J., Townsend, J., Cowles, A., Covington, J., Egaas, B., … Press, G. A. (2000). Normal brain development and aging: Quantitative analysis at in vivo MR imaging in healthy volunteers. *Radiology*, *216*(3), 672–682. https://doi.org/10.1148/radiology.216.3.r00au37672

De Vis, J. B., Zwanenburg, J. J., van der Kleij, L. A., Spijkerman, J. M., Biessels, G. J., Hendrikse, J., & Petersen, E. T. (2016). Cerebrospinal fluid volumetric MRI mapping as a simple measurement for evaluating brain atrophy. *European Radiology*, *26*(5), 1254–1262. https://doi.org/10.1007/s00330-015-3932-8

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., … Milham, M. P. (2014). The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, *19*(6), 659–667. https://doi.org/10.1038/mp.2013.78

Dubuisson, M.-P., & Jain, A. K. (2002). A modified Hausdorff distance for object matching. In *Proceedings of 12th International Conference on Pattern Recognition* (Vol. 1, pp. 566–568). Washington, DC: IEEE Computer Society Press.

Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., … Snaith, R. P. (2009). The Australian imaging, biomarkers and lifestyle (AIBL) study of aging: Methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, *21*(4), 672–687. https://doi.org/10.1017/S1041610209009405

Fjell, A. M., Westlye, L. T., Amlien, I., Espeseth, T., Reinvang, I., Raz, N., … Walhovd, K. B. (2009). High consistency of regional cortical thinning in aging across multiple samples. *Cerebral Cortex*, *19*(9), 2001–2012. https://doi.org/10.1093/cercor/bhn232

Fleischman, D., Berdahl, J. P., Zaydlarova, J., Stinnett, S., Fautsch, M. P., & Allingham, R. R. (2012). Cerebrospinal fluid pressure decreases with older age. *PLoS One*, *7*(12), e52664. https://doi.org/10.1371/journal.pone.0052664

François Chollet and contributors. (2015). keras. Retrieved from https://github.com/fchollet/keras

Frazier, J. A., Hodge, S. M., Breeze, J. L., Giuliano, A. J., Terry, J. E., Moore, C. M., … Makris, N. (2008). Diagnostic and sex effects on limbic volumes in early-onset bipolar disorder and schizophrenia. *Schizophrenia Bulletin*, *34*(1), 37–46. https://doi.org/10.1093/schbul/sbm120

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016* (Vol. 3, pp. 1651–1660). Retrieved from http://www.jmlr.org/proceedings/papers/v48/gal16.html

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: Mit Press. Retrieved from. http://www.deeplearningbook.org

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., … Margulies, D. S. (2015). NeuroVault.org: A web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in Neuroinformatics*, *9*, 8. https://doi.org/10.3389/fninf.2015.00008

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/J.NEUROIMAGE.2009.06.060

Hebert, L. E., Weuve, J., Scherr, P. A., & Evans, D. A. (2013). Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology*, *80*(19), 1778–1783. https://doi.org/10.1212/WNL.0b013e31828726f5

Heckemann, R. A., Hartkens, T., Leung, K. K., Zheng, Y., Hill, D. L. G., Hajnal, J. V, & Daniel, R. (2003). Information extraction from medical images: Developing an e–science application based on the Globus toolkit. In *Proceedings of UK e-Science all Hands Meeting*.

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., & Zhuowen, T. (2011). Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, *30*(9), 1617–1634. https://doi.org/10.1109/TMI.2011.2138152

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. Retrieved from http://arxiv.org/abs/1502.03167

Irimia, A., Torgerson, C. M., Goh, S.-Y. M., & Van Horn, J. D. (2015). Statistical estimation of physiological brain age as a descriptor of senescence rate during adulthood. *Brain Imaging and Behavior*, *9*(4), 678–689. https://doi.org/10.1007/s11682-014-9321-0

Iscan, Z., Jin, T. B., Kendrick, A., Szeglin, B., Lu, H., Trivedi, M., … DeLorenzo, C. (2015). Test-retest reliability of freesurfer measurements within and between sites: Effects of visual approval process. *Human Brain Mapping*, *36*(9), 3472–3485. https://doi.org/10.1002/hbm.22856

Ito, K., Fujimoto, R., Huang, T.-W., Chen, H.-T., Wu, K., Sato, K., … Aoki, T. (2018). Performance Evaluation of Age Estimation from T1-Weighted Images Using Brain Local Features and CNN. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 694–697). Piscataway, NJ: IEEE.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., … Weiner, M. W. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, *27*(4), 685–691. https://doi.org/10.1002/jmri.21049

Jäncke, L., Mérillat, S., Liem, F., & Hänggi, J. (2015). Brain size, sex, and the aging brain. *Human Brain Mapping*, *36*(1), 150–169. https://doi.org/10.1002/hbm.22619

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2015), 825–841. https://doi.org/10.1006/nimg.2002.1132

Kamnitsas, K., Chen, L., & Ledig, C. (2015). Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. Ischemic Stroke. Retrieved from http://www.isles-challenge.org/ISLES2015/pdf/20150930_ISLES2015_Proceedings.pdf#page=21

Kandel, B. M., Wolk, D. A., Gee, J. C., & Avants, B. (2013). Predicting cognitive data from medical images using sparse linear regression. *Information Processing in Medical Imaging*, *7917*, 86–97. https://doi.org/10.1007/978-3-642-38868-2_8

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* (pp. 5575–5585). Montreal, Canada: NIPS 2018.

Kennedy, D. N., Haselgrove, C., Hodge, S. M., Rane, P. S., Makris, N., & Frazier, J. A. (2012). CANDIShare: A resource for pediatric neuroimaging data. *Neuroinformatics*, *10*(3), 319–322. https://doi.org/10.1007/s12021-011-9133-y

Koen, J. D., & Yonelinas, A. P. (2014, September 15). The effects of healthy aging, amnestic mild cognitive impairment, and Alzheimer's disease on recollection and familiarity: A meta-analytic review. *Neuropsychology Review*, *24*, 332–354. https://doi.org/10.1007/s11065-014-9266-5

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, *51*(2), 181–207. https://doi.org/10.1023/A:1022859003006

Laird, A. R., Bzdok, D., Kurth, F., Fox, P. T., & Eickhoff, S. B. (2011). Activation likelihood estimation meta-analysis revisited. *NeuroImage*, *59*(3), 2349–2361. https://doi.org/10.1016/j.neuroimage.2011.09.017

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Retrieved from http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2323. https://doi.org/10.1109/5.726791

Ledig, C., Schuh, A., Guerrero, R., Heckemann, R. A., & Rueckert, D. (2018). Structural brain imaging in Alzheimer's disease and mild cognitive impairment: Biomarker analysis and shared morphometry database. *Scientific Reports*, *8*(1), 11258. https://doi.org/10.1038/s41598-018-29295-9

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., & Batra, D. (2015). Why M heads are better than one: Training a diverse ensemble of deep networks. *ArXiv*. Retrieved from http://arxiv.org/abs/1511.06314

Lemaitre, H., Goldman, A. L., Sambataro, F., Verchinski, B. A., Meyer-Lindenberg, A., Weinberger, D. R., & Mattay, V. S. (2012). Normal age-related brain morphometric changes: Nonuniformity across cortical thickness, surface area and gray matter volume? *Neurobiology of Aging*, *33*(3), 617.e1–617.e9. https://doi.org/10.1016/j.neurobiolaging.2010.07.013

Li, G., Liu, M., Sun, Q., Shen, D., & Wang, L. (2018). Early diagnosis of autism disease by multi-channel CNNs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11046 LNCS, pp. 303–309). Heidelberg, Germany: Springer Verlag. https://doi.org/10.1007/978-3-030-00919-9_35

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S. K., Huntenburg, J. M., … Margulies, D. S. (2016). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, *148*, 179–188. https://doi.org/10.1016/j.neuroimage.2016.11.005

Liu, W., Wei, D., Chen, Q., Yang, W., & Meng, J. (2017). Data descriptor: Longitudinal test–retest neuroimaging data from healthy young adults in Southwest China. *Scientific Data*, *4*, 1–9. https://doi.org/10.1038/sdata.2017.17

Lorenzi, M., Pennec, X., Frisoni, G. B., & Ayache, N. (2015). Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images. *Neurobiology of Aging*, *36*(S1), S42–S52. https://doi.org/10.1016/j.neurobiolaging.2014.07.046

Lorio, S., Kherif, F., Ruef, A., Melie-Garcia, L., Frackowiak, R., Ashburner, J., … Draganski, B. (2016). Neurobiological origin of spurious brain morphological changes: A quantitative MRI study. *Human Brain Mapping*, *37*(5), 1801–1815. https://doi.org/10.1002/hbm.23137

Marcus, D. S., Fotenos, A. F., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2010). Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. *Journal of*

Cognitive Neuroscience, *22*(12), 2677–2684. https://doi.org/10.1162/jocn.2009.21407

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., & Buckner, R. L. (2007). 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. https://doi.org/10.1162/jocn.2007.19.9.1498

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., … Taylor, P. (2011). The Parkinson progression marker initiative (PPMI). *Progress in Neurobiology*, *95*(4), 629–635. https://doi.org/10.1016/j.pneurobio.2011.09.005

Mayer, A. R., Ruhl, D., Merideth, F., Ling, J., Hanlon, F. M., Bustillo, J., & Cañive, J. (2013). Functional imaging of the hemodynamic sensory gating response in schizophrenia. *Human Brain Mapping*, *34*(9), 2302–2312. https://doi.org/10.1002/hbm.22065

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., & Lancaster, J. (1995). A probabilistic atlas of the human brain: Theory and rationale for its development. *NeuroImage*, *2*(2), 89–101. https://doi.org/10.1006/nimg.1995.1012

Mikhael, S. S., & Pernet, C. (2019). A controlled comparison of thickness, volume and surface areas from multiple cortical parcellation packages. *BMC Bioinformatics*, *20*(1), 55. https://doi.org/10.1186/s12859-019-2609-8

Monté-Rubio, G. C., Falcón, C., Pomarol-Clotet, E., & Ashburner, J. (2018). A comparison of various MRI feature types for characterizing whole brain anatomical differences using linear pattern recognition methods. *NeuroImage*, *178*, 753–768. https://doi.org/10.1016/j.neuroimage.2018.05.065

Mori, S., Wakana, S., van Zijl, P., & Nagae-Poetscher, L. (2005). *MRI atlas of human white matter*. Amsterdam: Elsevier.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. *Proceedings of the 27th International Conference on Machine Learning*, *3*, 807–814.

Neyshabur, B., Tomioka, R., & Srebro, N. (2015). In search of the real inductive bias: On the role of implicit regularization in deep learning. In *3rd International Conference on Learning Representations, ICLR 2015—Workshop Track Proceedings*. Retrieved from http://arxiv.org/abs/1412.6614

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*, *3* (3), e10.

Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, *35*(5), 1240–1251. https://doi.org/10.1109/TMI.2016.2538465

Pinel, P., Fauchereau, F., Moreno, A., Barbot, A., Lathrop, M., Zelenika, D., … Dehaene, S. (2012). Genetic variants of FOXP2 and KIAA0319/TTRAP/THEM2 locus are associated with altered brain activation in distinct language-related regions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(3), 817–825. https://doi.org/10.1523/JNEUROSCI.5996-10.2012

Poldrack, R. A., Congdon, E., Triplett, W., Gorgolewski, K. J., Karlsgodt, K. H., Mumford, J. A., … Bilder, R. (2016). A phenome-wide examination of neural and cognitive function. *BioRxiv*, 059733.

Qi, Q., Du, B., Zhuang, M., Huang, Y., & Ding, X. (2018). Age estimation from MR images via 3D convolutional neural network and densely connect. In *Lecture Notes in Computer Science* (Vol. 11307 LNCS, pp. 410–419). Cham: Springer.

Raz, N., Ghisletta, P., Rodrigue, K. M., Kennedy, K. M., & Lindenberger, U. (2010). Trajectories of brain aging in middle-aged and older adults: Regional and individual differences. *NeuroImage*, *51*(2), 501–511. https://doi.org/10.1016/J.NEUROIMAGE.2010.03.020

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249. https://doi.org/10.1002/widm.1249

Shamir, L., & Long, J. (2016). Quantitative machine learning analysis of brain MRI morphology throughout aging. *Current Aging Science*, *9*, 310–317. Retrieved from. http://www.ingentaconnect.com/contentone/ben/cas/2016/00000009/00000004/art00009

Simard, P. Y., Steinkraus, D., & Platt, J. (2003). Best practices for convolutional neural networks applied to visual document analysis. *In Proceedings of the Seventh International Conference on Document Analysis and Recognition-Volume 2* (p. 958). Washington, DC: IEEE Computer Society Press.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. Retrieved from http://arxiv.org/abs/1312.6034

Singh, S., Hoiem, D., & Forsyth, D. (2016). Swapout: Learning an ensemble of deep architectures. Retrieved from http://papers.nips.cc/paper/6205-swapout-learning-an-ensemble-of-deep-architectures

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. *ArXiv*. Retrieved from https://arxiv.org/abs/1706.03825

Sowell, E. R., Thompson, P. M., & Toga, A. W. (2004). *Mapping changes in the human cortex throughout the span of life. Neuroscientist*. Thousand Oaks, CA: Sage.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *In 3rd International Conference on Learning Representations, ICLR 2015—Workshop Track Proceedings*. Retrieved from http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a

Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., & Walhovd, K. B. (2014). Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: Regions of accelerating and decelerating change. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34* (25), 8488–8498. https://doi.org/10.1523/JNEUROSCI.0391-14.2014

Tanno, R., Worrall, D., Kaden, E., Ghosh, A., Grussu, F., Bizzi, A., … Alexander, D. C. (2019). Uncertainty quantification in deep learning for safer neuroimage enhancement. Retrieved from http://arxiv.org/abs/1907.13418

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* (pp. 2962–2971). Piscataway, NJ: IEEE. https://doi.org/10.1109/CVPR.2017.316

Valizadeh, S. A., Hänggi, J., Mérillat, S., & Jäncke, L. (2017). Age prediction on the basis of brain anatomical measures. *Human Brain Mapping*, *38* (2), 997–1008. https://doi.org/10.1002/hbm.23434

van der Walt, S., Yu, T., Gouillart, E., Yager, N., Nunez-Iglesias, J., Schönberger, J. L., … Warner, J. D. (2014). Scikit-image: Image processing in python. *PeerJ*, *2*, e453. https://doi.org/10.7717/peerj.453

Vanasse, T. J., Fox, P. M., Barron, D. S., Robertson, M., Eickhoff, S. B., Lancaster, J. L., & Fox, P. T. (2018). BrainMap VBM: An environment for structural meta-analysis. *Human Brain Mapping*, *39*(8), 3308–3325. https://doi.org/10.1002/hbm.24078

Vinke, E. J., de Groot, M., Venkatraghavan, V., Klein, S., Niessen, W. J., Ikram, M. A., & Vernooij, M. W. (2018). Trajectories of imaging markers in brain aging: The Rotterdam study. *Neurobiology of Aging*, *71*, 32–40. https://doi.org/10.1016/j.neurobiolaging.2018.07.001

Voevodskaya, O., Simmons, A., Nordenskjöld, R., Kullberg, J., Ahlström, H., Lind, L., … Alzheimer's Disease Neuroimaging Initiative, A. D. N. (2014). The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease. *Frontiers in Aging Neuroscience*, *6*, 264. https://doi.org/10.3389/fnagi.2014.00264

Wang, J., Knol, M. J., Tiulpin, A., Dubost, F., de Bruijne, M., Vernooij, M. W., … Roshchupkin, G. V. (2019). Gray matter age prediction as a biomarker for risk of dementia. *Proceedings of the National Academy of Sciences*, *116*(42), 21213–21218. https://doi.org/10.1073/pnas.1902376116

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., … Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *NeuroImage*, *45*(1), S173–S186. https://doi.org/10.1016/j.neuroimage.2008.10.055

Yang, C., Rangarajan, A., & Ranka, S. (2018). Global Model Interpretation via Recursive Partitioning. 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 1563–1570. https://doi.org/10.1103/PhysRevB.80.081304

Ziegler, G., Dahnke, R., Jäncke, L., Yotter, R. A., May, A., & Gaser, C. (2012). Brain structural trajectories over the adult lifespan. *Human Brain Mapping*, *33*(10), 2377–2389. https://doi.org/10.1002/hbm.21374

Zou, K. H., Warfield, S. K., Bharatha, A., Tempany, C. M. C., Kaus, M. R., Haker, S. J., … Kikinis, R. (2004). Statistical validation of image segmentation quality based on a spatial overlap index. *Academic Radiology*, *11*(2), 178–189. https://doi.org/10.1016/S1076-6332(03)00671-8

Zuo, X.-N., Anderson, J. S., Bellec, P., Birn, R. M., Biswal, B. B., Blautzik, J., … Song, X. W. (2014). An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data*, *1*, 140049. https://doi.org/10.1038/sdata.2014.49

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Levakov G, Rosenthal G, Shelef I, Raviv TR, Avidan G. From a deep learning model back to the brain—Identifying regional predictors and their relation to aging. *Hum Brain Mapp*. 2020;41:3235–3252. https://doi.org/10.1002/hbm.25011