DEEP LEARNING IN AGING NEUROSCIENCE

EDITED BY: Javier Ramírez, Juan Manuel Gorriz, Andres Ortiz, James H. Cole and Martin Dyrba

PUBLISHED IN: Frontiers in Neuroinformatics and Frontiers in Aging Neuroscience

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88966-281-4
DOI 10.3389/978-2-88966-281-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

DEEP LEARNING IN AGING NEUROSCIENCE

Topic Editors:
Javier Ramírez, University of Granada, Spain
Juan Manuel Gorriz, University of Granada, Spain
Andres Ortiz, University of Malaga, Spain
James H. Cole, University College London, United Kingdom
Martin Dyrba, Helmholtz Association of German Research Centers (HZ), Germany

Citation: Ramírez, J., Gorriz, J. M., Ortiz, A., Cole, J. H., Dyrba, M., eds. (2020). Deep Learning in Aging Neuroscience. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-281-4

Table of Contents

frontiers in Neuroinformatics

EDITORIAL
published: 26 October 2020
doi: 10.3389/fninf.2020.573974

Editorial: Deep Learning in Aging Neuroscience

Javier Ramírez[1], Juan M. Górriz[1], Andrés Ortiz[2], James H. Cole[3,4] and Martin Dyrba[5]*

[1] Department Signal Theory, Networking and Communications, University of Granada, Granada, Spain, [2] Department of Communications Engineering, University of Málaga, Málaga, Spain, [3] Department of Computer Science, Centre for Medical Image Computing, University College London, London, United Kingdom, [4] Dementia Research Centre, Institute of Neurology, University College London, London, United Kingdom, [5] German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany

Keywords: deep learning, convolutional neural networks, brain age estimation, neurodegenerative diseases, automated diagnosis, brain image segmentation

Editorial on the Research Topic

Deep Learning in Aging Neuroscience

1. INTRODUCTION

Deep learning (DL) has revolutionized the field of artificial intelligence by enabling computational models consisting of multiple processing layers to learn abstract representations of data (Hinton et al., 2006; Bengio et al., 2006). Conventional machine learning methods have been limited for decades by the need of expert knowledge to design sophisticated feature extraction algorithms in the process of transforming raw data into a suitable form for classification. In contrast, deep leaning methods, as representation-learning techniques, enable the learning model to be fed directly with raw data in order to discover the representations needed for classification (Krizhevsky et al., 2017; LeCun et al., 2015).

Currently, an intensive research effort is being devoted to the development of novel neuroimaging techniques to better understand the mechanisms of the central nervous system (CNS) and to early recognize age-related neural diseases (Payan and Montana, 2015; Sarraf and Tofighi, 2016; Martinez-Murcia et al., 2020; Martinez-Murcia et al., 2018, 2016) Ortiz et al.. The vast amount of data provided by large multicentre studies investigating new biomarkers for age-related neural diseases presents an opportunity for the development of more accurate deep learning models for early recognition of neurodegeneration as well as the characterization of the progressive course of neural disorders (Cole and Franke, 2017; Marzban et al., 2019; Segovia et al., 2018; Ortiz et al., 2016; Wang et al., 2018).

2. RESEARCH TOPIC CONTENT

The aim of this research topic "Deep Learning in Aging Neuroscience," published in *Frontiers in Aging Neuroscience* and *Frontiers in Neuroinformatics*, was to present the current state of the art in the theory and practice of deep learning computational modeling techniques in aging neuroscience with special emphasis on advancing our understanding of the mechanisms of CNS aging and age-related neural diseases. The research topic features 9 research articles. Most of the contributions examined disease progression and the relationships between different underlying pathological changes. Based on their contributions, the research articles were grouped into three main areas: brain age estimation (2 papers), automatic diagnosis of neurodegenerative diseases (5 papers), and brain image segmentation models (2 papers).

OPEN ACCESS

Edited and reviewed by:
Sean L. Hill,
Centre for Addiction and Mental Health, Canada

****Correspondence:***
Javier Ramírez
javierrp@ugr.es

Received: *22 June 2020*
Accepted: *16 September 2020*
Published: *26 October 2020*

Citation:
Ramírez J, Górriz JM, Ortiz A, Cole JH and Dyrba M (2020) Editorial: Deep Learning in Aging Neuroscience. Front. Neuroinform. 14:573974. doi: 10.3389/fninf.2020.573974

2.1. Brain Age Estimation

Predicting brain age based on structural and functional image data is still a challenging problem. Unveiling the normal aging of the brain is crucial in understanding the neural correlates of cognitive aging and neurodegenerative diseases such as Alzheimer's disease (AD).

Two papers of the research topic focused on brain age prediction by means of 18F-FDG brain metabolic topography data and structural T1-weighted MRI brain scans. Choi et al. proposed a deep learning model for predicting future brain metabolic topography by generating brain PET images. The generative model was based on a variational autoencoder (VAE). It used an 18F-FDG PET subject image and current age information as inputs to extract low-dimensional representation latent features that served as a basis for generation of PET image patterns corresponding to different brain ages. It was shown that, in spite of individual variability in age-related change, future regional metabolic changes were precisely predicted. The paper by Amoroso et al. presented an approach to predict brain aging based on structural T1-weighted MRI brain scans. They combined a complex network framework with deep learning strategies. Multiplex networks consisting of many layers were constructed, with each layer representing a single subject, the nodes being anatomical brain regions and the connections being derived from their pairwise similarities. A deep neural network processed nodal metrics, evaluating both the intensity and the uniformity of connections, to predict subjects' ages. The model yielded high accuracies and compared favorably with other state-of-the-art approaches.

2.2. Automatic Diagnosis of Neurodegenerative Diseases

This research topic also grouped different approaches for automatic diagnosis of neurodegenerative diseases based on deep learning classification models. Qiao et al. proposed a deep learning classification framework which performed multivariate data-driven feature extraction for automatic diagnosis of AD. The method was based on a three-level hierarchical partner matching independent component analysis (3LHPM-ICA) and Granger causality (GC) to infer the directional interaction between the independent components and to extract the effective connectivity features. Finally, a directed acyclic graph (DAG) neural network was used for classification. The proposed methodology was evaluated on a resting-state fMRI dataset consisting of 34 AD dementia patients and 34 normal controls (NCs) leading to a classification accuracy of 95.6%, with a sensitivity of 97.1% and a specificity of 94.1% with leave-one-out cross validation.

The paper by Qureshi et al. showed an evaluation of functional decline in AD dementia using three-dimensional convolutional neural networks (3D-CNN) and group ICA to model functional connectivity from resting-state fMRI data. The authors divided the dataset of AD patients into two groups based on dementia severity with respect to clinical dementia rating (CDR) scores: very mild to mild (CDR: 0.5–1) vs. moderate to severe (CDR: 2–3) dementia. Results reported a mean balanced classification accuracy of 92.3%, with specificity of 94.6% and sensitivity of 89.6%. In addition, medial frontal, sensorimotor, executive control, dorsal attention, and visual networks were found to be correlated with dementia severity.

Deep learning techniques showed improved classification performance in many medical imaging tasks including AD detection based on structural MRI data. However, these models are still perceived as being highly non-transparent and difficult to translate into clinical practice. The paper by Böhle et al. proposed layer-wise relevance propagation (LRP) to visualize CNN decisions for AD based on MRI data. This technique yields importance or relevance maps indicating how much each voxel is contributing to the final classification outcome, thus showing the potential of LRP to assist clinicians in interpreting neural network decisions.

The systematic review by Jo et al. presents recent studies using deep learning and neuroimaging data for diagnostic classification of AD. The authors included 16 research articles published between 2013 and 2018 and classified them according to deep learning algorithms and neuroimaging data types. Current state-of-the-art DL approaches yielded accuracies of up to 96.0% for AD dementia classification and 84.2% for the prediction of conversion from mild cognitive impairment (MCI) to dementia. The latter is of particular clinical relevance, as this could eventually lead to early identification of AD patients, enabling stratification for clinical trials and targeted interventions to delay dementia onset. However, the current accuracy of approximately 85% is likely too low for clinical adoption.

A deep learning approach for Parkinson's disease (PD) detection using isosurface-based features and convolutional neural networks was presented in Ortiz et al.. The authors proposed the use of isosurfaces as a solution to efficiently reduce the amount of data while keeping the most relevant information. Here, isosurfaces connect voxels above a specified intensity in a way similar to contour lines connecting points of equal elevation. These isosurfaces were then used to implement a classification system based on the CNN architectures LeNet and AlexNet. An average accuracy of 95.1% and AUC of 0.97 was achieved to differentiate PD patients and controls using 123I-Iofluopane (DaTSCAN) SPECT images. Finally, saliency maps of the last two-neuron layer were provided to determine which areas of the input brain images had a greater contribution to the predicted output class.

2.3. Brain Image Segmentation

Another important topic of research in the field of aging neuroscience is the development of image segmentation techniques for the assessment of disease progression. The paper by Jeong et al. focused on the segmentation of white matter hyperintensities (WMH) that appear as regions of abnormally high signal intensity on T2-weighted magnetic resonance image (MRI) sequences. This imaging marker has been identified as valuable for dementia and brain aging processes in age-related neuroscience. The authors developed and evaluated a saliency U-Net with irregularity age map (IAM) to decrease the U-Net architectural complexity without performance loss. Their so-called Dilated Saliency U-Net for WMH segmentation reduced

the training complexity of the original U-Net segmentation network and improved the Dice coefficient score to 0.56 with a sensitivity of 47.5%.

Segmentation of ischaemic stroke lesions remains a challenge in neuroimaging, especially when dealing with magnetic resonance perfusion imaging data. Deep learning CNN architectures developed to date reported low performance when segmenting ischaemic stroke lesions due to the lesion heterogeneity with respect to location, shape, size, image intensity and texture. The paper by Pérez Malla showed an evaluation of enhanced learning techniques for segmenting ischaemic stroke lesions in brain magnetic resonance perfusion images using a CNN scheme. In this way, data augmentation, transfer learning and post-processing techniques were evaluated for the segmentation of stroke lesions using the ISLES 2017 dataset, which contains expert annotated diffusion-weighted perfusion and diffusion brain MRI of 43 stroke patients. Among the experiments conducted, data augmentation combined with a binary hole filling procedure achieved the best results, improving the mean Dice score by 17% compared to the baseline model.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was partly supported by the MINECO/FEDER under TEC2015-64718-R, RTI2018-098913-B-100, PGC2018-098813-B-C32 projects and the General Secretariat for Universities, Research and Technology of the Junta de Andalucía under the A-TIC-117-UGR18 FEDER Andalucía project.

REFERENCES

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2006). Greedy layer-wise training of deep networks. In *Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS - 06* (Cambridge, MA: MIT Press), 153–160.

Cole, J. H., and Franke, K. (2017). Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends Neurosci.* 40, 681–690. doi: 10.1016/j.tins.2017.10.001

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Martinez-Murcia, F. J., Górriz, J. M., Ramírez, J., and Ortiz, A. (2016). A structural parametrization of the brain using hidden Markov models-based paths in Alzheimer's disease. *Int. J. Neural Syst.* 26:1650024. doi: 10.1142/S0129065716500246

Martinez-Murcia, F. J., Górriz, J. M., Ramírez, J., and Ortiz, A. (2018). Convolutional neural networks for neuroimaging in Parkinson's disease: is preprocessing needed? *Int. J. Neural Syst.* 28:1850035. doi: 10.1142/S0129065718500351

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J., Ramirez, J., and Castillo-Barnes, D. (2020). Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. *IEEE J. Biomed. Health Inform.* 24, 17–26. doi: 10.1109/JBHI.2019.2914970

Marzban, E. N., Teipel, S. J., Buerger, K., Fliessbach, K., Heneka, M. T., Kilimann, I., et al. (2019). P3-361: explainable convolutional networks and multimodal imaging data: the next step towards using artificial intelligence as diagnostic tool for early detection of Alzheimer's disease. *Alzheimers Dement.* 15(7S_Part_21), P1083–P1084. doi: 10.1016/j.jalz.2019.06.3394

Ortiz, A., Munilla, J., Górriz, J. M., Ramírez, J. (2016). Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. *Int. J. Neural Syst.* 26:1650025. doi: 10.1142/S0129065716500258

Payan, A., and Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506*.

Sarraf, S., and Tofighi, G. (2016). *DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI.* Cold Spring Harbor Laboratory. Available online at: https://www.biorxiv.org/content/early/2016/08/21/070441

Segovia, F., Górriz, J. M., Ramírez, J., Martinez-Murcia, F. J., and García-Pérez, M. (2018). Using deep neural networks along with dimensionality reduction techniques to assist the diagnosis of neurodegenerative disorders. *Logic J. IGPL* 26, 618–628. doi: 10.1093/jigpal/jzy026

Wang, S.-H., Phillips, P., Sui, Y., Liu, B., Yang, M., and Cheng, H. (2018). Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *J. Med. Syst.* 42:85. doi: 10.1007/s10916-018-0932-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 12 July 2018
doi: 10.3389/fnagi.2018.00212

Predicting Aging of Brain Metabolic Topography Using Variational Autoencoder

Hongyoon Choi[1], Hyejin Kang[1] and Dong Soo Lee[1,2,3]*, for the Alzheimer's Disease Neuroimaging Initiative[†]*

[1] Department of Nuclear Medicine, Seoul National University College of Medicine, Seoul, South Korea, [2] Department of Molecular Medicine and Biopharmaceutical Sciences, Graduate School of Convergence Science and Technology, Seoul National University, Seoul, South Korea, [3] Korea Brain Research Institute, Daegu, South Korea

OPEN ACCESS

Edited by:
Javier Ramírez, Universidad de Granada, Spain

Reviewed by:
Tonio Ball, Translational Neurotechnologie Labor, Albert-Ludwigs-Universität Freiburg, Germany
Nicola Amoroso, Università degli studi di Bari Aldo Moro, Italy

***Correspondence:**
Hongyoon Choi
chy1000@snu.ac.kr
Dong Soo Lee
dsl@plaza.snu.ac.kr

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Received: *20 March 2018*
Accepted: *22 June 2018*
Published: *12 July 2018*

Citation:
Choi H, Kang H and Lee DS for the Alzheimer's Disease Neuroimaging Initiative (2018) Predicting Aging of Brain Metabolic Topography Using Variational Autoencoder. Front. Aging Neurosci. 10:212. doi: 10.3389/fnagi.2018.00212

Predicting future brain topography can give insight into neural correlates of aging and neurodegeneration. Due to variability in the aging process, it has been challenging to precisely estimate brain topographical change according to aging. Here, we predict age-related brain metabolic change by generating future brain ^{18}F-Fluorodeoxyglucose PET. A cross-sectional PET dataset of cognitively normal subjects with different age was used to develop a generative model. The model generated PET images using age information and characteristic individual features. Predicted regional metabolic changes were correlated with the real changes obtained by follow-up data. This model was applied to produce a brain metabolism aging movie by generating PET at different ages. Normal population distribution of brain metabolic topography at each age was estimated as well. In addition, a generative model using APOE4 status as well as age as inputs revealed a significant effect of APOE4 status on age-related metabolic changes particularly in the calcarine, lingual cortex, hippocampus, and amygdala. It suggested APOE4 could be a factor affecting individual variability in age-related metabolic degeneration in normal elderly. This predictive model may not only be extended to understanding the cognitive aging process, but apply to the development of a preclinical biomarker for various brain disorders.

Keywords: brain metabolism, FDG PET, variational autoencoder, deep generative model, APOE4

INTRODUCTION

Understanding the normal aging change in the brain is essential to investigate neural correlates of cognitive aging and various neurodegenerative diseases including Alzheimer's disease (Jagust et al., 2009). In particular, the brain metabolism which can be measured by ^{18}F-fluorodeoxyglucose (FDG) PET has been regarded as a key biomarker for neurodegenerative disorders. Identifying brain metabolic topography associated with aging could give insight into the neural basis of age-related cognitive decline and help differentiate normal aging from neurodegenerative disorders.

Although the relationship between cerebral glucose metabolism and aging has been repeatedly studied, there has been controversy about which brain regions show significant age-related metabolic decline (Duara et al., 1984; Loessner et al., 1995; Moeller et al., 1996; Petit-Taboue et al., 1998; Yanase et al., 2005). Individual genetic background and healthy status as well as underlying

brain disease give rise to the individual variability in age-related metabolic change (Raz et al., 2005; Grady, 2012). Due to this variability, we have not been able to predict individual aged brain understandably. Instead of consideration of individual variability, previous studies have focused on the trend of overall aging changes using cross-sectional imaging data with statistical models such as linear regression. Even though this statistical analysis could provide overall brain metabolic changes, it was difficult to individually apply to estimate how far a given subject's brain metabolism is from the normal population at the same age. This individual evaluation of brain metabolism can be extended to the differentiation between normal and abnormal aging process. It requires normal population distribution database of all ages, however, it has been challenging to build a database of the population distribution of normal brain metabolism for each age from the limited cross-sectional data with subjects of various age distribution.

Here, we develop a model for predicting future brain metabolic topography by generating brain PET image. In this study, we utilize variational autoencoder (VAE), a type of unsupervised learning methods, which can generate images from some representations (VAE) (Kingma and Welling, 2013). We applied it to predicting FDG brain PET at different ages. Each FDG PET image combined with the subject's current age information was represented by low-dimensional features and then PET images corresponding different ages were generated. We also generated population distribution data of normal brain metabolic topography at different ages, which represented variability in individual metabolic activity at each age. As an application of our approach to discovering factors that potentially affect brain aging, we further investigated whether APOE4 status impacted on the age-related metabolic change by using a generative model that uses age and APOE4 information.

MATERIALS AND METHODS

Subjects

In this study, the data included subjects recruited in Alzheimer's Disease Neuroimaging Initiative (ADNI) with FDG PET images (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI recruited subjects from over 50 sites across the US and Canada. The primary purpose of ADNI has been to test whether serial imaging and biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see http://www.adni-info.org. Written informed consent to cognitive testing and neuroimaging prior to participation was obtained, approved by the institutional review boards of all participating institutions. Three hundred and ninety three cognitively normal subjects without Alzheimer's dementia or mild cognitive impairment performed baseline FDG PET (Age: 73.7 $\pm$ 5.9, range 56.1–90.1). These PET data and their age information were used for developing the model. All subjects underwent the clinical and cognitive assessment at the time of acquisition. APOE genotyping was performed on DNA samples obtained from blood. For detailed information on DNA sample preparation and genotyping, see http://www.adni-info.org. For 393 subjects, 113 (28.8%) were APOE4 carriers and 280 (71.2%) were APOE4 non-carriers.

FDG PET Preparation

All the PET images were downloaded from ADNI database. FDG PET images were acquired 30 to 60 min and the images were averaged across the time frames and standardized to have same voxel size ($1.5 \times 1.5 \times 1.5$ mm). PET images were acquired in the 57 sites participating in ADNI, scanner-specific smoothing was additionally applied (Jagust et al., 2015). PET images were spatially normalized to the Montreal Neurological Institute (MNI) space using statistical parametric mapping (SPM8, www.fil.ion.ucl.ac.uk/spm). Each PET image was divided by mean FDG uptake of the cerebellum for normalization.

Variational Autoencoder for PET Volumes

We utilized VAE model to generate virtual PET data according to age information. VAE-based PET image generation is summarized in **Figure 1A**. VAE is a type of unsupervised learning methods which could represent the high-dimensional data to low-dimensional features. The major strength of the VAE is to generate virtual data from latent features. VAE consisted of two components, encoder and generator. The encoder reduces the dimension of data by compressing them to latent features and the generator produces the data from any values of latent features. The generator of VAE is a probabilistic generator which assumes that the data were generated from some conditional distribution and an unobserved variable z in latent space. Thus, the probabilistic generator can be defined by $p_\theta(x|z)$. θ represents the parameters of generator. The posterior distribution $p_\theta(z|x)$ can be obtained by prior distribution $p(z)$, $p_\theta(z|x) \sim p(z)\, p_\theta(x|z)$. Variational Bayes learns both parameters, $p_\theta(x|z)$ and an approximation $q_\phi(z|x)$ to the intractable true posterior $p_\theta(z|x)$. This is achieved by the loss function,

$$L(\phi,\theta) = -E_{z\sim q_\phi(z|x)}(\log p_\theta(x|z)) + KL(q_\phi(z|x) \parallel p_\theta(z))$$

where KL is Kullback-Leibler divergence between the learnt latent distribution and the prior distribution $p_\theta(z)$, acting as a regularization term (Kingma and Welling, 2013). The first term represents reconstruction loss of autoencoder.

In this study, we applied VAE with age information to generate PET image, so used VAE conditioning on another description of the data, y (i.e., age information). This model is aimed to generate data from the conditional distribution as well as latent features z. Thus, the probabilistic generator and the encoder can be defined by $p_\theta(x|y, z)$ and $q_\phi(z|x, y)$, respectively. The loss function is changed to,

$$L(\phi,\theta) = -E_{z\sim q_\phi(z|x,y)}(\log p_\theta(x|y,z)) + KL(q_\phi(z|x,y) \parallel p_\theta(z))$$

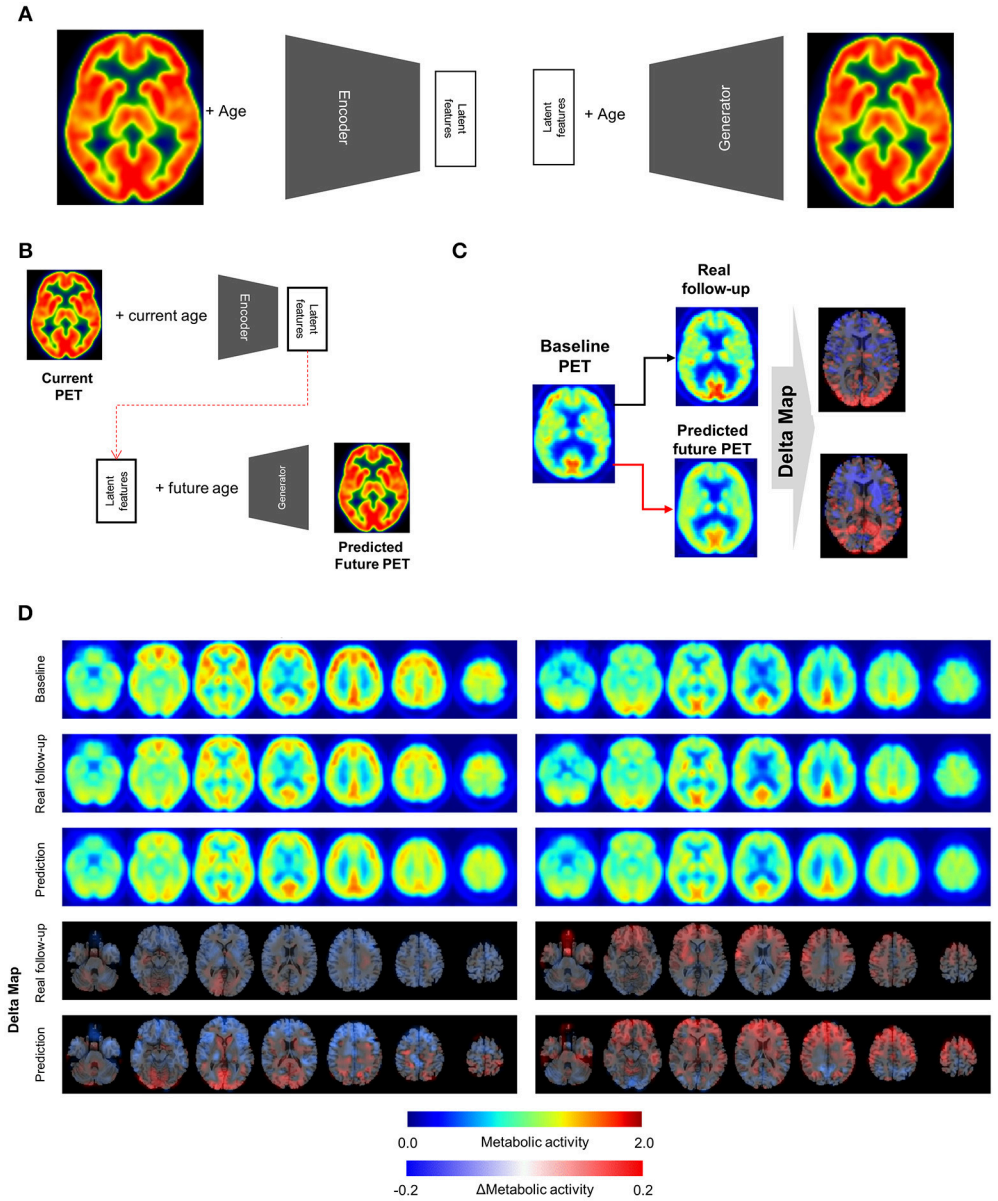

FIGURE 1 | Metabolic change prediction by generating future brain PET. **(A)** VAE model which consists of encoder and generator was trained by PET images of cognitively normal subjects. The encoder represents input PET images to 10 latent features. The generator generates virtual PET image from any values of latent features and age information. **(B)** The VAE-based model could generate future brain PET individually using baseline PET image. A subject's brain PET was encoded into latent features. We hypothesized that these latent features were unchanged across age. Future brain PET was generated by entering future age and the latent features. **(C)** Predicted individually generated PET was compared with real follow-up data. For comparison, delta maps obtained by subtracting baseline from prediction or follow-up images were generated. **(D)** Representative cases follow-up PET and individually predicted PET. According to the follow-up data, there was comparable individual variability in metabolic change. A subject showed globally decreased metabolism (left) while another subject showed increased metabolism in the frontotemporal cortex (right). Predicted future PET could also reflect the individual variability.

To train VAE, data X and age information y were encoded into parameters in a latent features Z, and decoder network reconstructs data from the latent features and y assuming latent features have normal distributions around encoded feature z. In practice, generator input was resampled by the encoded latent features z assuming normal distribution:$z_{resampled} = z_{encoder} + z_{sd} \times \varepsilon$, where ε represents a random variable (Kingma and Welling, 2013).

Network Architecture and Training

To encode 3-dimensional PET volume, we used multiple 3D convolutional layers for encoding. Specific parameters for network architecture are summarized in Supplementary Figure 1.

After the multiple convolutional and pooling layers, 3D feature volumes are changed to 1-dimensional features. These features are merged by age information of each subject and additionally connected to hidden layers and, finally, connected to 10 latent features. Accordingly, initial PET volume with $79 \times 95 \times 68$ matrix is compressed into 10-dimensional features. Conversely, the generator consists of convolutional and upsampling layers. Upsampling simply repeats each dimension of the data. Input variables of the generator include 10 latent features and age information. The generator decodes these inputs to PET volume.

This conditional VAE model was trained by gradient descent algorithm (Adadelta) (Zeiler, 2012) and took 50 epochs for the training. The VAE was implemented using a deep learning library, Keras (ver. 1.2.2) with Theano (ver. 0.9.0) backend (Bastien et al., 2012). Ten percentage of all PET data were used for the validation set to determine epoch number and hyperparameters for the neural network architecture.

Estimation of Metabolic Activity in Brain Regions

The regional metabolic activity of brain regions was obtained using predefined volume-of-interests (VOI), automated anatomical labeling (AAL) template. As all PET images were spatially normalized to MNI template, mean metabolic activity value of each brain region was simply obtained by masking specific brain region.

Prediction of Future PET and Comparison With Follow-up PET

Four-year follow-up FDG brain PET scans were obtained in 26 cognitively normal subjects who underwent baseline PET scans. Five-year follow-up FDG brain PET scans were acquired in 11 cognitively normal subjects. Longitudinal change in brain metabolism was evaluated in these subjects. Using baseline PET images of the subjects and age, we generated future PET images. To generate individual future PET image, firstly, baseline PET image was represented into latent features using the encoder. We hypothesized that these latent features were unchanged regardless of subject's age. Ten latent features of a subject and future age (i.e., baseline age + 4 or 5) were used for the generator. We compared real follow-up PET and predicted PET by using delta maps. To measure similarity between predicted and real metabolic changes, voxelwise correlation coefficient was calculated. Similarity measurements were individually obtained. We statistically tested whether other variables including baseline age, gender, APOE4 status, Mini-Mental State Examination (MMSE) and follow-up diagnosis affected the prediction of metabolic changes. The similarity measurements, correlation coefficients, of the group according to the APOE4 status, gender and follow-up diagnosis were statistically compared using independent *t*-test. They were correlated with continuous variables (age and MMSE) using Pearson correlation. We also additionally evaluated the overall accuracy of predicted image using mean absolute percentage error (MAPE). MAPE between predicted and real follow-up PET image was calculated for each subject.

In addition, overall predicted and real regional changes were calculated by AAL map. The overall regional metabolic change was calculated by mean value across all subjects. The correlation between regional metabolic changes of predicted and real follow-up PET across brain regions was tested by Pearson correlation. For visualizing the similarity between predicted and real metabolic changes, Bland-Altman plots were drawn. Ninety percentage confidence interval for error of predicted regional metabolic change was calculated.

Generation of Age-Related Metabolic Change Movie

The overall age-related metabolic change pattern was evaluated by the generator model. Firstly, PET data of all subjects were represented by 10 latent features using the encoder. The mean feature values were entered into the generator with different age information between the age of 50 and 100. Thus, we could obtain representative PET image of each age. To visualize age-related metabolic change, we generate subtraction map. Generated PET images with different age were subtracted by a representative brain PET generated by age of 50. These subtraction maps were also visualized by an animation.

Population Distribution of Regional Metabolic Activity at Each Age

We estimated population distribution of regional metabolic activity by resampling generated PET images. Ten latent features were randomly resampled assuming each latent feature has normal distribution. Mean and standard deviation of each latent feature were determined by the feature values of all subjects. One Thousand resampled brain PET images were generated and regional metabolic activity was obtained. Population distribution of metabolic activity of each region was drawn by histograms and age-related changes with confidence intervals were drawn.

Metabolic Topography According to Latent Features

To assess the relationship between latent features and brain metabolic patterns, brain PET images were generated by changing values of the latent features. Mean values of latent features were used for generating PET except for two features for estimating effects on brain metabolism. These two features were changed from −2.0 to 2.0 and generated virtual PET images for plotting.

Variability in Age-Related Metabolic Change According to the APOE4 Status

To evaluate age-related metabolic change patterns according to the APOE4 status, another VAE model was trained. Conditional VAE with age and APOE4 status information was used, so, conditional variable, y, includes age and APOE4 status as different dimensions. The training process and network architectures were same with conditional VAE with age only.

The overall age-related metabolic change patterns according to APOE4 status was evaluated as population distribution estimation. Randomly resampled latent features and different age values were entered into the generator with each APOE4 status respectively. PET images of each age and APOE4 status were generated and regional metabolic activity was obtained by

predefined regions. Population distribution of regional metabolic activity was estimated for APOE4 carriers and non-carriers. To find statistically different regions, we calculated the difference between regional metabolic activity generated by APOE4 carriers and non-carriers using randomly resampled latent feature values. To define statistical significance, p-values were computed by the distribution of the difference. The null hypothesis was that the regional metabolic difference between APOE4 carriers and non-carriers is 0. Thus, the statistical significance could be directly calculated by the proportion of the generated samples where the difference was lower or higher than 0. Brain regions with different metabolic activity were found at each age. The difference with uncorrected p-value < 0.05 was regarded as significant brain regions which show different metabolism according to APOE4 status.

RESULTS

Prediction of Future Brain Metabolic Change

The VAE-based model was designed to represent FDG PET images and corresponding subjects' age to latent features (**Figure 1A**). The posterior part of this model, the generator component, could produce PET images from any values of the latent features and age information.

To generate future brain PET images, we firstly obtained latent features of a subject's baseline PET image using the encoder. We assumed that these were not changed according to aging as characteristic individual features. The features of a subject were entered into the generator with any age, which could generate the subject's virtual brain PET at different ages (**Figure 1B**). The model was tested by cognitively healthy subjects who underwent both baseline and follow-up PET. The predicted metabolic change was compared with corresponding real metabolic change computed by follow-up PET data. Each predicted future brain PET and real follow-up PET was subtracted from corresponding baseline PET for the comparison (**Figure 1C**). As a result, delta maps, the future brain PET subtracted by the baseline, obtained from real follow-up PET showed individual variability. Corresponding predicted future brain PET also showed those variable patterns (**Figure 1D**). A subject showed prominently decreased metabolism in the cerebral cortices, while another showed relatively increased metabolism in the frontal cortex (**Figure 1D**). The delta map obtained by real follow-up was positively correlated with that obtained by prediction (Supplementary Figure 2).

To compare predicted future brain PET and real follow-up PET quantitatively, mean metabolic changes of 116 predefined brain regions across all subjects were calculated. Averaged predicted changes in regional metabolism was significantly correlated with the real changes obtained by real follow-up data ($r = 0.59$, $p < 0.001$ and $r = 0.59$, $p < 0.001$ for 4-year and 5-year follow-up, respectively; **Figures 2A,B**). Bland-Altman plots showed the difference between predicted and real regional metabolic activities (**Figures 2C,D**). The 95% confidence interval of the prediction error of regional metabolic activity was -0.027 to 0.027 for 4-year follow-up and -0.027 to 0.048 for 5-year follow-up. In addition, individually predicted and real metabolic changes were compared. To show how individual prediction of metabolic change was similar to the real change, voxelwise correlations of individual delta maps obtained by follow-up and prediction were calculated. We could find a trend of high correlation between the two delta maps of the same subject, even though the prediction of metabolic change was failed in some subjects (Supplementary Figure 3). The similarity between predicted and real metabolic change was not significantly affected by subjects' age, gender, follow-up diagnosis, APOE4, and baseline MMSE. As a global measurement of overall accuracy for predicting future brain PET, we obtained MAPE by comparing predicted PET with real follow-up PET. MAPE was 7.8 ± 2.1 and $8.3 \pm 1.5\%$ for 4-year and 5-year follow-up, respectively. Notably, MAPE calculated by baseline PET and reconstructed PET using VAE was $6.6 \pm 1.4\%$ (**Figure 2E**).

Generating Overall Brain Metabolism Aging Movie

We applied our model to the assessment of overall regional metabolic changes. To investigate overall patterns of age-related brain metabolism, representative brain images were generated by using different age and mean value of each latent feature across all subjects (**Figure 3A**). The representative FDG brain PET generated from the age of 50 to 90 is presented in **Figure 3B**. To visualize the age-related change definitely, the generated FDG PET with different age was subtracted by the generated PET of the age of 50 (**Figures 3C,D,** Supplementary Figure 4). As we generated brain metabolic topography at all ages, overall age-related patterns were also visualized by movies (Supplementary Movies 1, 2).

Figure 3D showed that age-related metabolism decline was mainly found in the cingulate cortex. Using predefined brain regions of interests, the metabolic activity of each brain region was extracted according to aging (**Figure 3E**). Red dotted lines represent estimated metabolic decline using the generated PET obtained by entering mean latent features. Solid lines represent real metabolic decline obtained by 4-year (Blue) and 5-year (Green) follow-up data (**Figure 3E**). The curves estimated by the VAE model explained that overall metabolic decline with aging was non-linear. Approximately before 75, the age-related metabolic decline was steep in the posterior cingulate and caudate and then the decline became slower after 75.

Distribution of Regional Metabolic Activity at Each Age

Most brain imaging data including our subjects consist of imaging with various ages. Thus, it has been challenging to obtain population distribution of normal brain at each age. Randomly resampled latent features could generate population distribution of regional brain metabolic activity for all ages (**Figure 4A**). Generated brain PET data from resampled latent features provide the variety of regional metabolic activity. Histograms of each brain region at different ages were drawn (**Figure 4B**). As aforementioned representative brain metabolic changes, histograms of posterior cingulate and caudate showed a trend of left shifting according to aging. Distribution of overall aging patterns of regional metabolism was also exhibited

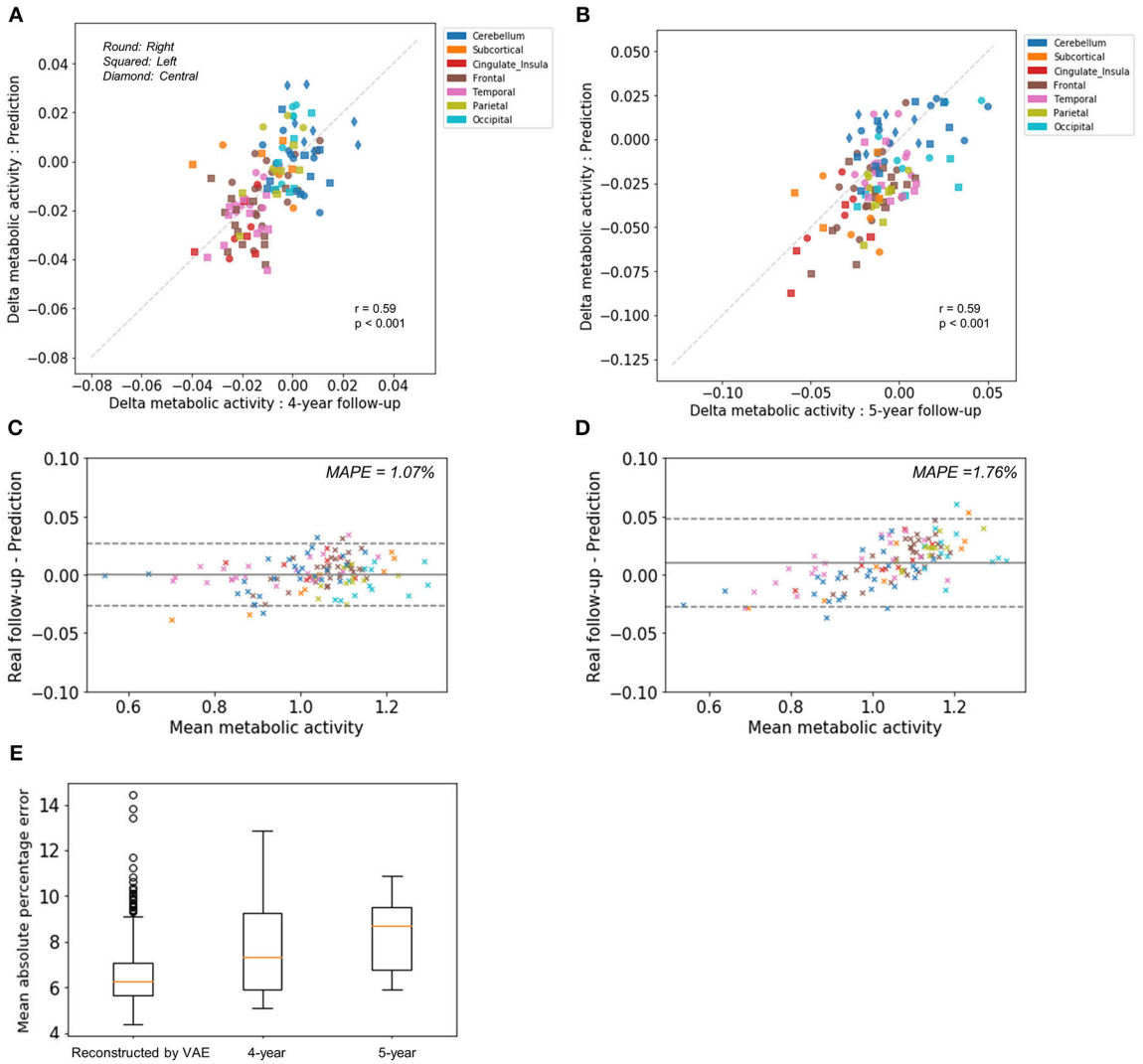

FIGURE 2 | Comparison of predicted metabolic change with real follow-up data**.** Regional metabolic change from baseline was averaged across subjects for predicted and follow-up data. Averaged predicted and real changes across the brain regions were significantly correlated for 4-year follow-up images ($r = 0.59$, $p < 0.001$) **(A)** and 5-year follow-up images ($r = 0.59$, $p < 0.001$) **(B)**. Bland-Altman plots were drawn for the comparison of predicted and real regional metabolic activity for 4-year **(C)** and 5-year PET images **(D)**. The 95% confidence interval of the error of predicted regional metabolism was −0.027 to 0.027 for 4-year follow-up and −0.027 to 0.048 for 5-year follow-up. Mean absolute percentage error (MAPE) was 1.07% for 4-year follow-up and 1.76% for 5-year follow-up. **(E)** As a global measurement of accuracy for predicting future brain PET, MAPE was 7.8 ± 2.1% for 4-year follow-up and 8.3 ±1.5% for 5-year follow-up. MAPE calculated by baseline PET and output of VAE with baseline age was 6.6 ± 1.4%.

(**Figure 4C,** Supplementary Figure 5). Dotted lines represent 95% confidence intervals of regional metabolic activity.

We found individual variability in regional brain metabolism at different ages. The individual variability was determined by the distribution of latent features. To show how each latent feature affects brain metabolism, PET images were generated by changing latent features. Brain metabolic patterns were changed according to latent features as shown in **Figure 5**. As an example, increased feature 1 was associated with decreased brain metabolism in the posterior temporal and occipital cortices and increased feature 2 was associated with increased frontal metabolism.

APOE4 Status and Age-Related Metabolic Change

Because clinical variables affect age-related metabolic change and its variability, we further investigated whether APOE4 status impacts on metabolic changing patterns. Another VAE model was trained using two conditions, age and APOE4 status (**Figure 6A**). This model can generate virtual brain PET images according to the age and APOE4 status. Thus, age-related metabolic change according to APOE4 can be estimated by inputting APOE4 positive and negative states, respectively (**Figure 6A**). We identified that APOE4 could affect the variability of age-related metabolic change. The FDG PET

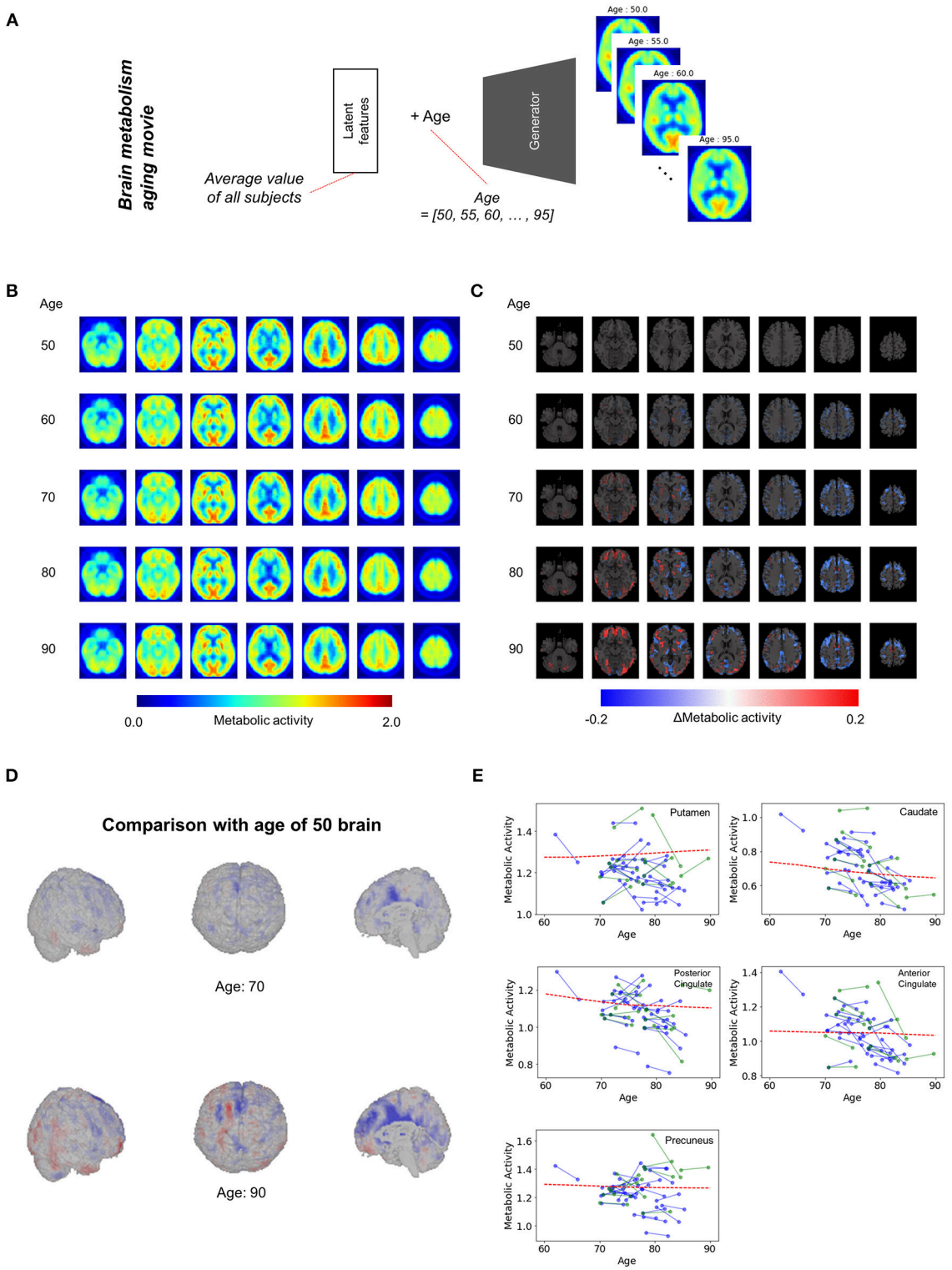

FIGURE 3 | Overall brain metabolism aging movie by generating representative PET of each age. **(A)** Using VAE-based model, representative FDG PET images of different age were generated to identify overall age-related metabolic pattern. Mean latent feature values across all trained subjects were entered into the generator for representative PET images. **(B)** Using mean latent features, representative PET images were generated according to aging. **(C)** Compared with the representative PET of age of 50, subtraction images were generated. **(D)** Surface visualization of the subtraction map revealed that age-related decline was mainly found in the cingulate cortex. **(E)** Age-related metabolic change in specific brain regions was plotted. Solid lines represent real metabolic change data for 4-year follow-up (blue) and 5-year follow-up (green). Red dotted lines represent regional metabolic changes estimated by virtually generated PET images.

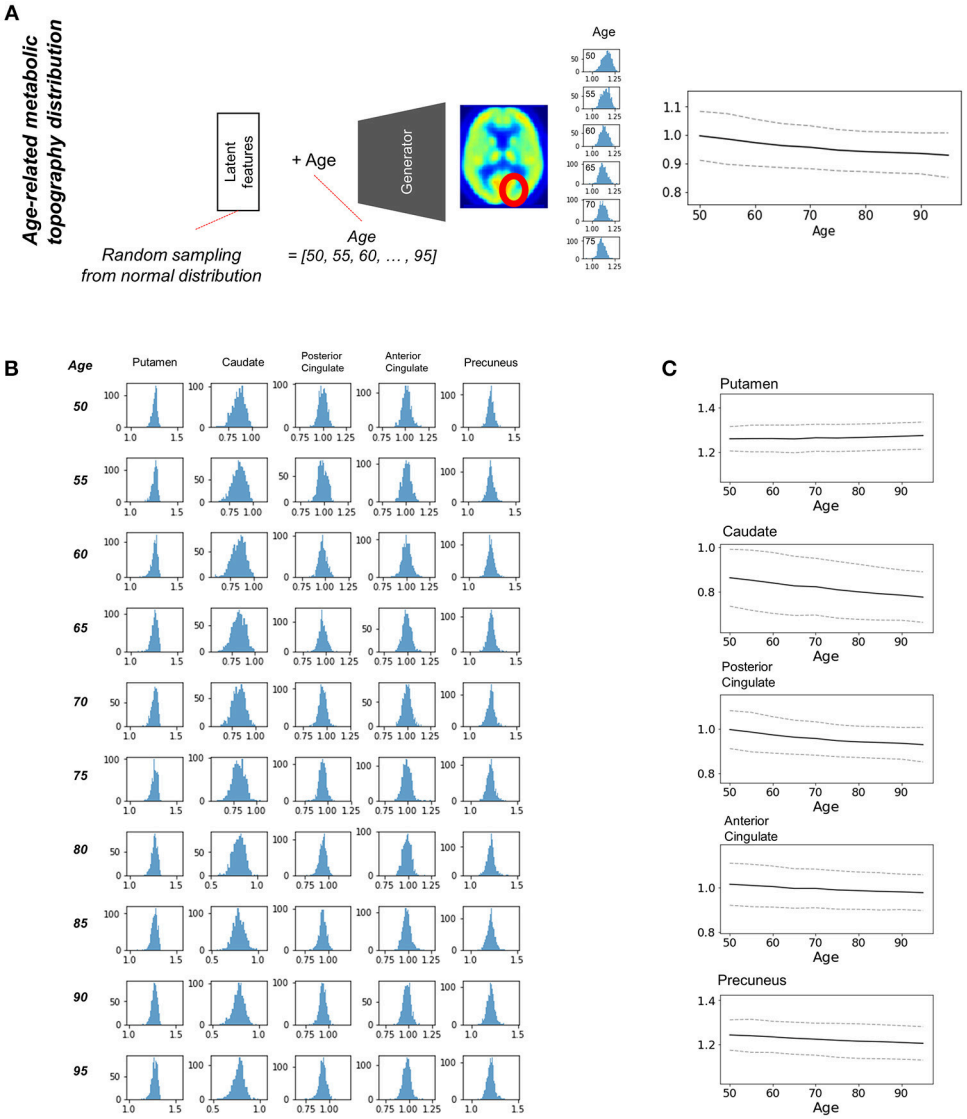

FIGURE 4 | Estimating population distribution of brain metabolism at each age**. (A)** Population distribution of brain metabolic topography was estimated by resampling latent features. Generated brain PET was repeatedly generated by random latent feature values sampled from the normal distribution. Distribution of regional metabolism was estimated for all ages. **(B)** Histograms of distribution of metabolic activity were drawn for putamen, caudate, posterior cingulate, anterior cingulate, and precuneus at different ages. **(C)** Confidence intervals of metabolic changes could be estimated by the distribution. Dotted lines represent 95% confidence interval of regional metabolic activity.

images generated by average latent features and APOE4 positive and negative status at different ages were subtracted by the generated PET of the age of 50 (**Figure 6B**). We found that metabolic decline in occipital lobe was faster in APOE4 carriers. Distribution of regional metabolism according to APOE4 status was estimated (**Figure 6C**). Using distribution of the metabolic difference between brain metabolism generated by APOE4 status, the significance of the difference in regional metabolism was estimated (Supplementary Figure 6). The regional metabolic activity of the calcarine and lingual cortex was significantly higher in APOE4 carrier than APOE4 non-carrier before 60, while that of the hippocampus and amygdala was significantly lower in APOE4 carrier at 50 (**Figure 6D**). The regional metabolic activity of posterior cingulate, precuneus and caudate, where the rapid age-related metabolic decline was found, did not show a significant difference in accordance with APOE4 status. Metabolic change in APOE4 carriers and non-carriers of all brain regions was represented with 95% confidence intervals (Supplementary Figure 7).

DISCUSSION

In this study, we predicted aging of brain metabolic topography by using a generative model. Brain metabolic changes are

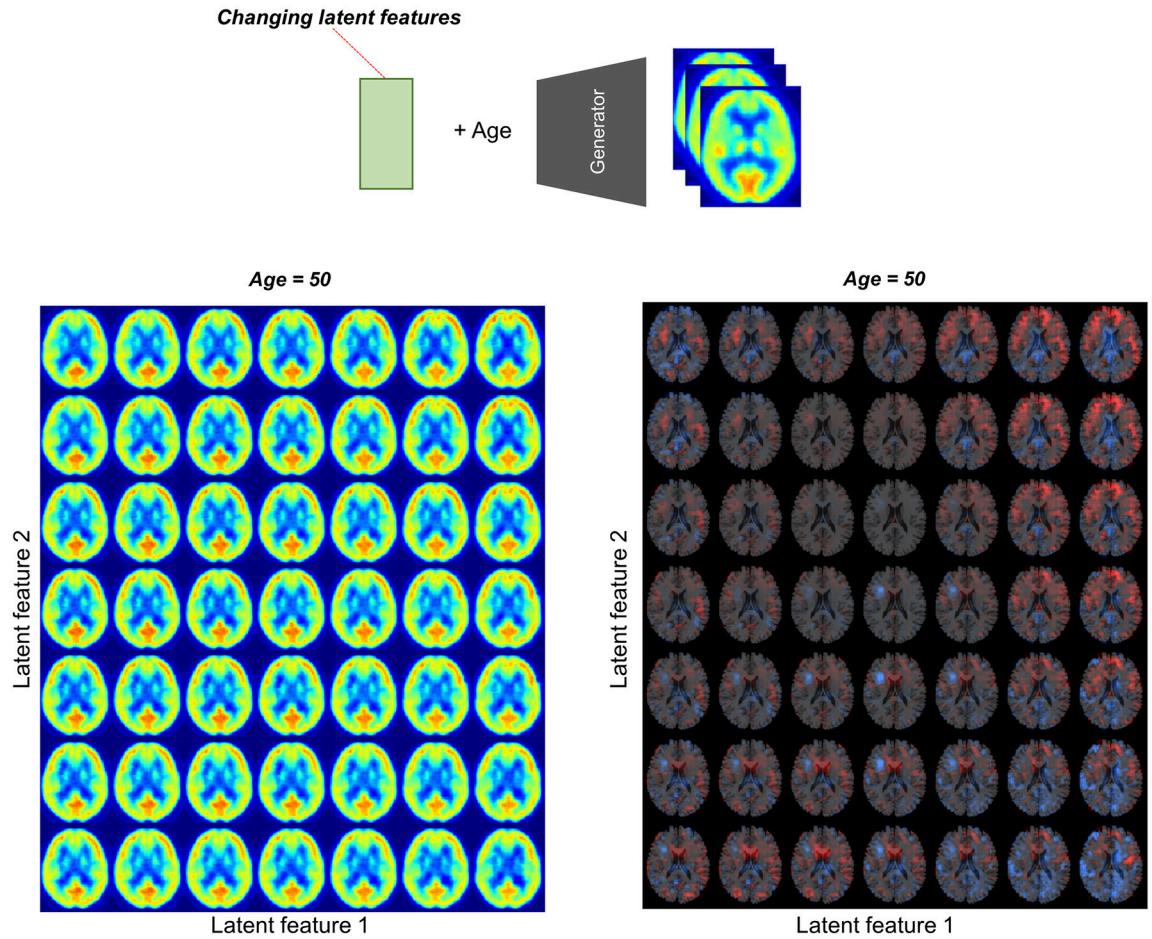

FIGURE 5 | Brain metabolic topography according to latent features**.** As the encoder of VAE compressed PET image into 10 latent features, variability in brain metabolism is determined by these 10 features. To assess metabolic patterns determined by latent features, brain PET images were generated according to different latent feature values. An example of the two latent features, increased first latent feature (x-axis) was associated with decreased metabolism in posterior temporal and occipital cortices. Increased second latent feature (y-axis) was associated with increased metabolism in the frontotemporal cortices at age of 50.

highly variable as aging process and cognitive changes are affected by several individual factors. Our model aimed at generating PET images according to the age trained by cross-sectional PET image data combined with different ages. The model could provide predicted future metabolic decline and was validated by real follow-up data. Our results estimate population distribution of normal brain metabolism at each age. This approach was extended to investigate the effect of APOE4 status on the variability of regional brain metabolism at different ages.

Our generative model could find population distribution of brain metabolic topography for each age as well as predict age-related metabolic change. Cognitive aging and the age-related functional decrease are accompanied by increased individual variability (Ylikoski et al., 1999). This individual variability is affected by several factors including life experience, genetic backgrounds, and susceptibility to neuropathology (Shammi et al., 1998). Furthermore, cognitive variability in individuals across time tends to occur mainly after the age of 60 (Wilson et al., 2002). Increased individual variability in aging has been supported by several functional neuroimaging studies (Glisky et al., 2001; D'esposito et al., 2003; Burzynska et al., 2015). Nonetheless, age-related brain metabolism change has been briefly estimated by observing an overall correlation between age and metabolism (Duara et al., 1984; Loessner et al., 1995; Moeller et al., 1996; Petit-Taboue et al., 1998; Yanase et al., 2005). This previous approach could not consider individual variability in age-related metabolism (Ylikoski et al., 1999; Knopman et al., 2014). Furthermore, it has been difficult to estimate age-dependent normal population distribution of brain image data as the data consist of subjects with different ages. A conventional linear regression model based on overall metabolic changes estimated by all baseline scans only estimated same decline patterns for all subjects by calculating a voxelwise linear regression based on the population.

According to our model, the individual variability of brain metabolism was represented by the latent features. The latent

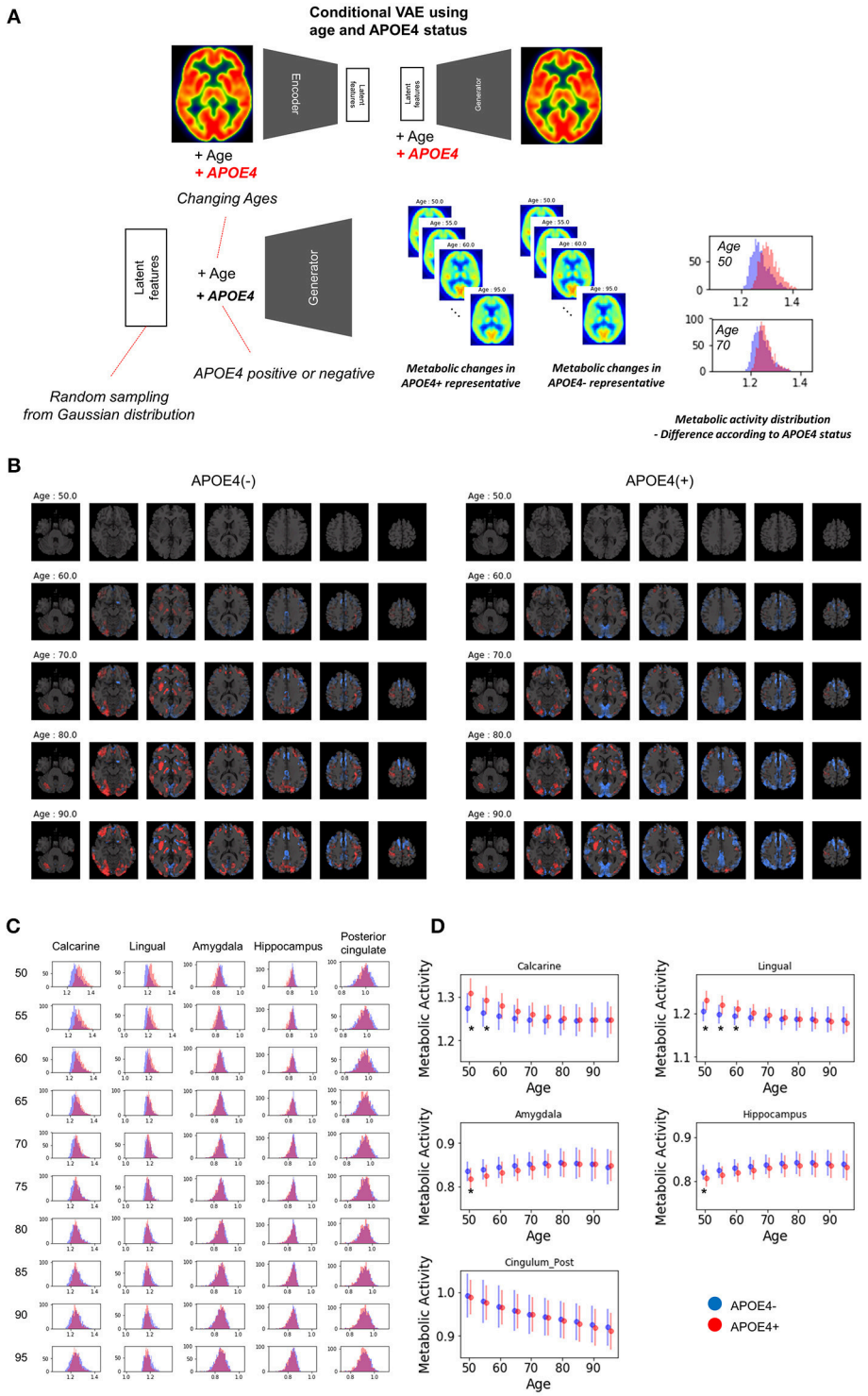

FIGURE 6 | APOE4 status and age-related brain metabolic change**. (A)** We investigated whether APOE4 status affect age-related metabolic change patterns. A conditional generative model was developed using APOE4 status as well as age. PET images according to different ages were generated for APOE4 carrier and non-carrier, respectively. Resampled features provide distribution difference of brain metabolic topography between APOE4 carriers and non-carriers. **(B)** Delta maps were generated by subtracting 50-year-old generated images. Metabolic decline was relatively faster in occipital regions of APOE4 carrier. **(C)** Histograms of regional metabolic activity were drawn for APOE4 carriers and non-carriers. Before 60, metabolism of calcarine, lingual cortex, hippocampus, and amygdala was different according to APOE4 status. **(D)** Age-related regional metabolic activity changes were plotted. Red dots represented APOE4 carriers and blue dots represented APOE4 non-carriers. Bars represented standard deviations calculated by the distribution. Non-parametric testing revealed the statistical significance. Asterisks represent uncorrected $p < 0.05$.

features determine age-related metabolism patterns because the generator used only the latent features and age as inputs. We could indirectly know whether the VAE model uses age information for generating PET images. Firstly, the generator of VAE model uses age information in that different metabolism distributions are shown by different age inputs and same latent feature values. Furthermore, we compared our VAE model with another VAE model without age information, which encodes latent features regardless of age and generates PET from latent features without age inputs. As a result, the VAE model without age information extracts more age-dependent latent features than the VAE model with age information (Supplementary Figure 8). It suggests that the VAE model without age information is prone to extract age-dependent image features with unsupervised manner because age-related changes largely contribute to the variability in brain metabolism. On the other hand, our model extracts individual characteristic image features relatively independent of age by using age information for the encoder. Each latent feature represented specific metabolic topography patterns which could be indirectly identified by generating images according to different feature values as shown in **Figure 5**. In this regard, random resampling of the latent features generated variable brain metabolic topography, which could be used for estimating population distribution. Our result, population distribution of brain metabolism at each age can be applied to quantitatively define regional abnormality in individuals. Using this distribution, we can define how far a given individual brain PET is from the normal population. Thus, this distribution may help to develop quantitative biomarker which represents abnormal aging process of individual brain metabolism.

Our model could predict regional patterns of individual future brain metabolic change, while future prediction of metabolic change was incorrect in quite a few cases. As shown in Supplementary Figure 3, the predicted delta maps were not correlated with real delta maps in individuals at right-lower portions of the matrix. Individual age-related changes measured by PET could be the sum of biologic metabolic change and statistical random errors in FDG PET. The statistical variability in brain metabolism could affect prediction accuracy of metabolic changes. Nonetheless, overall regional metabolic changes obtained by the prediction were highly correlated with those of real follow-up data as shown in **Figure 2**. That was because VAE eventually extracted age-associated metabolic topography patterns from overall variation of brain metabolism in the training samples. In other words, because of the high variability in age-related brain metabolic changes, VAE-based model generated future brain PET image by approximating global age-related patterns of training samples. It is closely related to the limitation of VAE which tends to generate averaged and blurry images and lack of variety in generated images (Dosovitskiy and Brox, 2016). Notably, though predicted overall regional metabolic changes in 5-year follow-up were significantly correlated with real follow-up data, they tend to underestimation in the regions with high metabolism and overestimation in those with low metabolism (**Figure 2D**). MAPE of 5-year follow-up was higher than 4-year follow-up as well as reconstructed images, which suggested the prediction accuracy could be affected by follow-up intervals. It could be due to long follow-up interval which could cause more non-aging factors affecting brain metabolism. Not only aging but several cognitive, healthy, and nutritional factors affect brain metabolic patterns (Belanger et al., 2011; Cunnane et al., 2011). Because of the multiple factors affecting brain metabolism, accurate individual prediction, particularly for long-term prediction, is substantially difficult. In this study, we simply assumed that other factors of future brain PET except age are unchanged. As multiple factors could determine metabolic topography, the generative model with multiple conditions such as cognitive score may improve future PET prediction.

Combination of another generative model such as generative adversarial model may improve the prediction accuracy (Goodfellow et al., 2014). Briefly, the generative adversarial model is another generative model using two networks, generator and discriminator. The generator is trained to synthesize images from latent features which cannot be discriminated from the training data, while discriminator is trained to discriminate real images from generated images. This type of model also can be combined with conditions such as aging information. The generative adversarial model can generate more realistic images compared with VAE, however, according to our experiments, 3-dimensional PET images were hard to generate using it. A further modification will be required to train the model and to generate more accurate future brain images. In addition, parameters including the number of latent features, model architectures and optimization methods could be modified to obtain better results. Although we tested several models, methods to develop optimized neural network architectures will be required as a future work.

Population distribution of metabolic topography revealed that APOE4 carriers showed higher metabolism in the calcarine and lingual cortex, while lower metabolism in the hippocampus and amygdala before 55. The difference in these regions was not found after 60, which suggested that age-related metabolic changes of these regions were greater in APOE4 carriers than non-carriers. The relationship between APOE4 and brain metabolism in normal elderly has been investigated in previous studies as well (Oh et al., 2014). The regions which showed difference metabolism in accordance with APOE4 status were partly different as the previous study showed that metabolic decline was faster in composite region-of-interests including posterior cingulate, precuneus, and lateral parietal cortices (Oh et al., 2014). Besides, another study using functional MRI showed APOE4 status affected the differentiation of functional networks including hippocampal and visual networks though they used different modality (Trachtenberg et al., 2012). Structural MRI study showed that APOE4 carriers tended to have thicker cortex in temporooccipital areas and a steeper age-related decline in cortical thickness (Espeseth et al., 2008). Although the regions related to APOE4 were partly different according to the studies, our result supports APOE4 carriers could affect functional brain aging patterns. Additionally, by estimating population distribution, we could identify regional

metabolic difference at all ages. Our approach can be extended to the investigation of the association between other clinical variables and age-related changes. It can eventually help find the factors that determine the individual variability in aging.

To our knowledge, this is the first report that applies a generative model to estimate aging of high dimensional medical data. As an extended application of our approach, PET data according to interpretable features, such as sex and cognitive scores, can be generated by using conditional VAE which aimed at synthesizing virtual data from the conditional distribution (Kingma et al., 2014; Sohn et al., 2015). This conditional generative model can be used for various problems in neuroimaging analyses. For example, the model may be used for predicting several task-specific functional brain images from a single image data. Virtual task-related brain images can be predicted by inputting tasks as conditional inputs of VAE model. Furthermore, this approach would improve conventional statistical voxelwise analyses of neuroimaging data. An important limitation of the voxelwise analysis is the presence of multiple covariates (Friston et al., 1994; Petersson et al., 1999). So far, covariates such as subject's age and brain volume have been handled as nuisance variables using general linear model. Instead, virtual neuroimaging data in same conditions can be generated by this approach. For instance, we can compare brain images of different groups by generating virtual data with controlled covariates such as same age and brain volume.

As a deep generative model may be able to precisely predict high dimensional data, the future application will be extended to various medical implications. Recently, generative models have been used in various biomedical fields as well as neuroimaging data. A generative model was applied to generating novel molecular fingerprints as an artificial intelligence drug discovery framework (Kadurin et al., 2017). As a recently developed application to medical image processing, a generative model was used for automatic lesion segmentation (Alex et al., 2017).

In our study, we predict aging of metabolic topography by generating PET images. In spite of individual variability in age-related change, future regional metabolic changes were precisely predicted. Population distribution of normal brain metabolism at different ages was estimated. It revealed that regional metabolic decline was different according to the APOE4 status. This brain metabolic change prediction method can provide a plausible explanation of individual variability in cognitive aging. Furthermore, we expect that this approach will be extended to the development of a preclinical biomarker for several neurodegenerative disorders as well as defining abnormal brain aging.

AUTHOR CONTRIBUTIONS

HC and DL designed the study. HC generated the model. HC and HK performed statistical analysis. All authors interpreted the data, wrote, and approved the manuscript.

FUNDING

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIP) (No. 2017M3C7A1048079). This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) (No.2017R1A5A1015626). This research was also supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI14C0466), and funded by the Ministry of Health & Welfare, Republic of Korea (HI14C3344), and funded by the Ministry of Health & Welfare, Republic of Korea (HI14C1277), and the Technology Innovation Program (10052749).

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.;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 Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

ACKNOWLEDGMENTS

In this study, the data included subjects recruited in Alzheimer's Disease Neuroimaging Initiative (ADNI) with FDG PET images (http://adni.loni.usc.edu).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2018.00212/full#supplementary-material

REFERENCES

Alex, V., Kp, M. S., Chennamsetty, S. S., and Krishnamurthi, G. (2017). "Generative adversarial networks for brain lesion detection," in *SPIE Medical Imaging*: *International Society for Optics and Photonics* (Orlando, FL).

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., et al., (2012). Theano: new features and speed improvements. *arXiv preprint arXiv:*1211.5590.

Bélanger, M., Allaman, I., and Magistretti, P. J. (2011). Brain energy metabolism: focus on astrocyte-neuron metabolic cooperation. *Cell Metab.* 14, 724–738. doi: 10.1016/j.cmet.2011.08.016

Burzynska, A. Z., Wong, C. N., Voss, M. W., Cooke, G. E., Mcauley, E., and Kramer, A. F. (2015). White matter integrity supports BOLD signal variability and cognitive performance in the aging human brain. *PLoS ONE* 10:e0120315. doi: 10.1371/journal.pone.0120315

Cunnane, S., Nugent, S., Roy, M., Courchesne-Loyer, A., Croteau, E., Tremblay, S., et al., (2011). Brain fuel metabolism, aging, and Alzheimer's disease. *Nutrition* 27, 3–20. doi: 10.1016/j.nut.2010.07.021

D'esposito, M., Deouell, L. Y., and Gazzaley, A. (2003). Alterations in the BOLD fMRI signal with ageing and disease: a challenge for neuroimaging. *Nat. Rev. Neurosci.* 4, 863–872. doi: 10.1038/nrn1246

Dosovitskiy, A., and Brox, T. (2016). "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems* (Barcelona), 658–666.

Duara, R., Grady, C., Haxby, J., Ingvar, D., Sokoloff, L., Margolin, R. A., et al., (1984). Human brain glucose utilization and cognitive function in relation to age. *Ann. Neurol.* 16, 703–713. doi: 10.1002/ana.410160613

Espeseth, T., Westlye, L. T., Fjell, A. M., Walhovd, K. B., Rootwelt, H., and Reinvang, I. (2008). Accelerated age-related cortical thinning in healthy carriers of apolipoprotein E epsilon 4. *Neurobiol. Aging* 29, 329–340. doi: 10.1016/j.neurobiolaging.2006.10.030

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* 2, 189–210. doi: 10.1002/hbm.460020402

Glisky, E. L., Rubin, S. R., and Davidson, P. S. (2001). Source memory in older adults: an encoding or retrieval problem? *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1131–1146. doi: 10.1037/0278-7393.27.5.1131

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Montréal, CA), 2672–2680.

Grady, C. (2012). The cognitive neuroscience of ageing. *Nat. Rev. Neurosci.* 13, 491–505. doi: 10.1038/nrn3256

Jagust, W. J., Landau, S. M., Koeppe, R. A., Reiman, E. M., Chen, K., Mathis, C. A., et al. (2015). The Alzheimer's Disease Neuroimaging Initiative 2 PET Core: 2015. *Alzheimers Dement.* 11, 757–771. doi: 10.1016/j.jalz.2015.05.001

Jagust, W. J., Landau, S. M., Shaw, L. M., Trojanowski, J. Q., Koeppe, R. A., Reiman, E. M., et al., (2009). Relationships between biomarkers in aging and dementia. *Neurology* 73, 1193–1199. doi: 10.1212/WNL.0b013e3181bc010c

Kadurin, A., Aliper, A., Kazennov, A., Mamoshina, P., Vanhaelen, Q., Khrabrov, K., et al. (2017). The cornucopia of meaningful leads: applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* 8, 10883–10890. doi: 10.18632/oncotarget.14073

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). "Semi-supervised learning with deep generative models," in *Advances in Neural Information Processing Systems* (Montréal, CA), 3581–3589.

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:*1312.6114.

Knopman, D. S., Jack, C. R., Wiste, H. J., Lundt, E. S., Weigand, S. D., Vemuri, P., et al., (2014). 18F-fluorodeoxyglucose positron emission tomography, aging, and apolipoprotein E genotype in cognitively normal persons. *Neurobiol. Aging* 35, 2096–2106. doi: 10.1016/j.neurobiolaging.2014.03.006

Loessner, A., Alavi, A., Lewandrowski, K. U., Mozley, D., Souder, E., and Gur, R. E. (1995). Regional cerebral function determined by FDG-PET in healthy volunteers: normal patterns and changes with age. *J. Nucl. Med.* 36, 1141–1149.

Moeller, J. R., Ishikawa, T., Dhawan, V., Spetsieris, P., Mandel, F., Alexander, G. E., et al. (1996). The metabolic topography of normal aging. *J. Cereb. Blood Flow Metab.* 16, 385–398. doi: 10.1097/00004647-199605000-00005

Oh, H., Habeck, C., Madison, C., and Jagust, W. (2014). Covarying alterations in Abeta deposition, glucose metabolism, and gray matter volume in cognitively normal elderly. *Hum. Brain Mapp.* 35, 297–308. doi: 10.1002/hbm.22173

Petersson, K. M., Nichols, T. E., Poline, J. B., and Holmes, A. P. (1999). Statistical limitations in functional neuroimaging. I. Non-inferential methods and statistical models. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 1239–1260. doi: 10.1098/rstb.1999.0477

Petit-Taboué M. C., Landeau, B., Desson, J. F., Desgranges, B., and Baron, J. C. (1998). Effects of healthy aging on the regional cerebral metabolic rate of glucose assessed with statistical parametric mapping. *Neuroimage* 7, 176–184. doi: 10.1006/nimg.1997.0318

Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., et al. (2005). Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cereb. Cortex* 15, 1676–1689. doi: 10.1093/cercor/bhi044

Shammi, P., Bosman, E., and Stuss, D. T. (1998). Aging, Neuropsychology, and Cognition *Aging Variability Perform.* 5, 1–13. doi: 10.1076/anec.5.1.1.23

Sohn, K., Lee, H., and Yan, X. (2015). "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems* (Montréal, CA), 3483–3491.

Trachtenberg, A. J., Filippini, N., Ebmeier, K. P., Smith, S. M., Karpe, F., and Mackay, C. E. (2012). The effects of APOE on the functional architecture of the resting brain. *Neuroimage* 59, 565–572. doi: 10.1016/j.neuroimage.2011.07.059

Wilson, R. S., Beckett, L. A., Barnes, L. L., Schneider, J. A., Bach, J., Evans, D. A., et al., (2002). Individual differences in rates of change in cognitive abilities of older persons. *Psychol. Aging* 17, 179–193. doi: 10.1037/0882-7974.17.2.179

Yanase, D., Matsunari, I., Yajima, K., Chen, W., Fujikawa, A., Nishimura, S., et al., (2005). Brain FDG PET study of normal aging in Japanese: effect of atrophy correction. *Eur. J. Nucl. Med. Mol. Imaging* 32, 794–805. doi: 10.1007/s00259-005-1767-2

Ylikoski, R., Ylikoski, A., Keskivaara, P., Tilvis, R., Sulkava, R., and Erkinjuntti, T. (1999). Heterogeneity of cognitive profiles in aging: successful aging, normal aging, and individuals at risk for cognitive decline. *Eur. J. Neurol.* 6, 645–652. doi: 10.1046/j.1468-1331.1999.660645.x

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:*1212.5701

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 17 December 2018
doi: 10.3389/fnagi.2018.00417

Multivariate Deep Learning Classification of Alzheimer's Disease Based on Hierarchical Partner Matching Independent Component Analysis

Jianping Qiao[1]†, Yingru Lv[2]†, Chongfeng Cao[3], Zhishun Wang[4]* and Anning Li[5]**

[1]Shandong Province Key Laboratory of Medical Physics and Image Processing Technology, Institute of Data Science and Technology, School of Physics and Electronics, Shandong Normal University, Jinan, China, [2]Department of Radiology, Huashan Hospital, Fudan University, Shanghai, China, [3]Department of Emergency, Jinan Central Hospital Affiliated to Shandong University, Jinan, China, [4]Department of Psychiatry, Columbia University, New York, NY, United States, [5]Department of Radiology, Qilu Hospital of Shandong University, Jinan, China

Machine learning and pattern recognition have been widely investigated in order to look for the biomarkers of Alzheimer's disease (AD). However, most existing methods extract features by seed-based correlation, which not only requires prior information but also ignores the relationship between resting state functional magnetic resonance imaging (rs-fMRI) voxels. In this study, we proposed a deep learning classification framework with multivariate data-driven based feature extraction for automatic diagnosis of AD. Specifically, a three-level hierarchical partner matching independent components analysis (3LHPM-ICA) approach was proposed first in order to address the issues in spatial individual ICA, including the uncertainty of the numbers of components, the randomness of initial values, and the correspondence of ICs of multiple subjects, resulting in stable and reliable ICs which were applied as the intrinsic brain functional connectivity (FC) features. Second, Granger causality (GC) was utilized to infer directional interaction between the ICs that were identified by the 3LHPM-ICA method and extract the effective connectivity features. Finally, a deep learning classification framework was developed to distinguish AD from controls by fusing the functional and effective connectivities. A resting state fMRI dataset containing 34 AD patients and 34 normal controls (NCs) was applied to the multivariate deep learning platform, leading to a classification accuracy of 95.59%, with a sensitivity of 97.06% and a specificity of 94.12% with leave-one-out cross validation (LOOCV). The experimental results demonstrated that the measures of neural connectivities of ICA and GC followed by deep learning classification represented the most powerful methods of distinguishing AD clinical data from NCs, and these aberrant brain connectivities might serve as robust brain biomarkers for AD. This approach also allows for expansion of the methodology to classify other psychiatric disorders.

Keywords: Alzheimer's disease, independent component analysis, granger causality, brain network, deep learning

OPEN ACCESS

Edited by:
Javier Ramírez, University of Granada, Spain

Reviewed by:
Ivan Sahumbaiev, Kyiv Polytechnic Institute, Ukraine
Stavros I. Dimitriadis, Cardiff University School of Medicine, United Kingdom

***Correspondence:**
Jianping Qiao
jpqiao@sdu.edu.cn
Zhishun Wang
wangz@nyspi.columbia.edu
Anning Li
anningli00@163.com

†*These authors have contributed equally to this work*

Received: *10 September 2018*
Accepted: *03 December 2018*
Published: *17 December 2018*

Citation:
Qiao J, Lv Y, Cao C, Wang Z and Li A (2018) Multivariate Deep Learning Classification of Alzheimer's Disease Based on Hierarchical Partner Matching Independent Component Analysis. Front. Aging Neurosci. 10:417. doi: 10.3389/fnagi.2018.00417

INTRODUCTION

Alzheimer's disease (AD) is a chronic neurodegenerative disease characterized by cognitive and intellectual deficits that are serious enough to interfere with daily life. It usually starts slowly and worsens over time by destroying brain cells, leading to memory loss, problems performing familiar tasks, vision problems, thinking, reasoning, and personality changes (Burns and Iliffe, 2009; Querfurth and LaFerla, 2010). Gradually, bodily functions are lost, ultimately leading to death (Alzheimer's Association, 2011). With the aging of the world population, AD has become a serious problem to the health the elderly people and a huge burden to the healthcare system. Nowadays, AD can only be slowed down and delayed by drugs, and effective treatment remains elusive (Jack et al., 2008). The diagnosis of AD is usually based on cognitive impairments relating to daily activities or positive physiopathologic markers of AD, such as an abnormal level of amyloid beta and/or tau in the cerebrospinal fluid (Dubois et al., 2014). Therefore, it is of great interest to develop objective biomarkers of AD patients with the help of neuroimaging studies in order to assist AD clinical diagnosis and monitor the efficacy of treatment.

Brain imaging technology, combined with advanced signal processing approaches, has been actively applied to investigate the underlying biological or neurological mechanisms and to discover differences between AD patients and normal controls (NCs) for AD diagnosis or prognosis (Mirzaei et al., 2016). Positron emission tomography (PET) accessed the pathophysiologic markers of AD as reductions of glucose metabolism in the parietal, posterior cingulate and temporal brain regions of AD patients (Diehl et al., 2004). Additionally, high resolution structural magnetic resonance imaging (sMRI) studies have shown that neuroimaging measurements included cortical thickness (Thompson et al., 2004; Lerch et al., 2008; Desikan et al., 2009; Dickerson et al., 2009), gray matter density (Dai et al., 2012; Liu M. et al., 2015; Liu et al., 2016), hippocampal volume and shape (Colliot et al., 2008; Fan et al., 2008; Hua et al., 2008; Chupin et al., 2009; Tsao et al., 2017). Histogram characteristics of regions of interest (ROIs) in the whole brain (Magnin et al., 2009) could be investigated as brain features for the classification between AD and NC. Furthermore, the measures of diffusion tensor imaging (DTI) such as fractional anisotropy (FA) and mean diffusivity (MD), which indicated white matter (WM) fiber tract integrity, have been reported to discriminate AD from NC (Dyrba et al., 2013). Another study reported that the WM tracts connecting brain regions defined by 41 Brodmann areas were reconstructed as the brain connectivity network and the graphs of the connectivity matrices were described as feature vectors for the classification of AD (Ebadi et al., 2017). Moreover, the absolute and relative spectral power, distribution of spectral power, and measures of spatial synchronization were calculated from recordings of the electroencephalography (EEG) by following classification models for the clinical diagnosis of AD (Lehmann et al., 2007). The lagged linear connectivity of predefined ROIs was also used as an EEG marker of AD (Babiloni et al., 2016; Triggiani et al., 2017).

Besides, resting state functional MRI (rs-fMRI) combined with machine learning has played an important role in identifying biomarkers of AD. Various classification features of AD have been detected in previous studies, such as the amplitude of low frequency fluctuations (Dai et al., 2012) or hippocampal correlation of low frequency components (Li et al., 2002), regional homogeneity (Dai et al., 2012), functional correlation strength of 90 ROIs in terms of the automated anatomical labeling (AAL) atlas (Dai et al., 2012), whole-brain (Chen et al., 2011; Ju et al., 2017) or selected regional (Wang K. et al., 2006) functional correlation connectivity matrices based on AAL or other atlas (Khazaee et al., 2016), covariance connectivity matrices (Challis et al., 2015), and graph-theoretical measures (Dyrba et al., 2015; Khazaee et al., 2015, 2017). However, most of the existing studies focus on seed-based correlation analysis which needed a prior (such as atlas) and ignored the relationship between voxels of brain images. The performance of the seed-based correlation methods may be unstable due to the different seeds or atlas as well as the error of the registration processing (Wang et al., 2009; Zalesky et al., 2010; Craddock et al., 2012). Therefore, as a multivariate data-driven based method, independent component analysis (ICA) was investigated to extract features for automatic classification of AD in the study, which could identify the underlying data structure by counting for the relationship between voxels and without need of prior information.

ICA has been widely applied for analyzing neuroimaging data (Calhoun et al., 2009) and acknowledged as one of the two most commonly used methods in functional connectivity (FC) studies (Zhang and Raichle, 2010). At present, there are two kinds of ICA methods applied to fMRI: individual ICA and group ICA. Previous studies have demonstrated that the AD patients displayed lower FC within the default mode network (DMN) identified by spatial individual ICA (Toussaint et al., 2014) or group ICA (Binnewijzend et al., 2012). A recent study reported that the FC matrices obtained by group ICA and the graph properties can be applied for the classification of AD (de Vos et al., 2018). However, compared with group ICA, the specificity of the individuals can be preserved better in the individual ICA method because a single temporally concatenated data set of all subjects is decomposed into ICs in group ICA. This leads to the possibility that the obtained ICs may not be maximally spatially independent for single subjects and degrades the precision of the identified functional brain network. Therefore, this study focuses on the individual ICA in order to extract the distinguishable features and predict the individuals with AD. However, there are still some problems in individual ICA method. First, the output order of ICs is uncertain, leading to the difficult establishment of the correspondence between the ICs or functional networks of multiple subjects. Second, the number of components must be defined before ICA is performed. Various brain functional networks might be obtained when the specified number is different. Lastly, the FC patterns resulting from multiple implementations of the same ICA algorithm on the same fMRI data may be inconsistent because of the randomness of the initial value in the ICA algorithm.

To address the issues mentioned above, we proposed a three-level hierarchical partner matching ICA (3LHPM-ICA) approach, which could identify the stable and reproducible ICs across multiple individuals. Then the extracted FC features were fused with the effective connectivity matrices computed by Granger causality (GC). Finally, the two-dimensional feature matrices were entered into the deep learning classifier to distinguish AD from NC. The aim of the current study was to detect the underlying fMRI data structure and biomarkers of AD with the multivariate data-driven based feature extraction and deep learning platform by counting for the relationship between voxels without needing prior information.

MATERIALS AND METHODS

Participants

Thirty-four participants with mild AD (17 females, 17 males, mean age 68.64 ± 9.85 years, education 11.47 ± 3.49 years) were recruited from a memory outpatient clinic at the Huashan Hospital of Fudan University. Thirty-four age-matched NCs (13 females, 21 males, mean age 65.55 ± 8.98 years, education 11.31 ± 3.75 years) were recruited by public advertisement to take part in the study. All AD participants fulfilled the following clinical criteria: the National Institute of Neurological and Communicative Disorders and Stroke/Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA; McKhann et al., 1984) criteria for AD, Mini Mental State Examination (MMSE) scores between 19 and 23 (inclusive), Clinical Dementia Rating (CDR) scores (Morris, 1993) of 1.0, Hachinski Ischemic Scale (HIS) scores less than 4.0 for the exclusion of vascular dementia and mixed dementia, and there were not any structural abnormalities other than atrophy in MRI scans. A standard diagnostic examination that included physical and neurological examination, medical history taking, extensive neuropsychological assessments and screening laboratory tests, was implemented for all patients. The mean MMSE score of AD group in this study was 21.50 ± 1.61. All NC subjects had normal neurological examinations, with a CDR score of 0 and independently functioning community membership with no history of neurological or psychiatric disorders, cognitive complaints, brain damage or psychoactive medication. All participants were right-handed with ten or more years of education. This study was carried out in accordance with the recommendations of NINCDS-ADRDA, the Institutional Review Board of Huashan Hospital of Fudan University with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Institutional Review Board of Huashan Hospital of Fudan University.

Image Acquisition

Imaging was performed on a Siemens Verio 3.0 Tesla MRI scanner (Siemens, Erlangen, Germany). The head of each participant was snugly fixed by using foam pads to reduce head movements and scanner noise. Participants were instructed to rest with their eyes closed but not to fall asleep during scanning. Resting state fMRI data were acquired using a T2*-weighted echoplanar imaging (EPI) with blood oxygen level dependent (BOLD) contrast pulse sequence. Thirty-three contiguous axial slices were acquired along the anterior commissure-posterior commissure (AC-PC) plane. The acquisition parameters were as follows: matrix = 64 × 64, field of view (FOV) = 20 cm, repetition time (TR) = 2,000 ms, echo time (TE) = 35 ms, voxel size = 3.0 × 3.0 × 4.0 mm^3, flip angle = 90°, slice thickness = 4 mm. The sequence took 6 min and 40 s, resulting in a total of 200 volumes.

Image Analysis

Preprocessing

All preprocessing steps of the resting state fMRI images were performed with SPM12 (Welcome Department of Imaging Neuroscience, London, United Kingdom) implemented in MATLAB. The functional scans were slice time corrected for the interleaved acquisition, spatially realigned to the first scan to correct for head movements, normalized to the Montreal Neurological Institute (MNI) coordinate system and spatially smoothed with an isotropic 8 mm full-width at half-maximum (FWHM) Gaussian kernel.

Functional Connectivity Analysis Based on 3LHPM-ICA

In this study, a 3LHPM-ICA approach was proposed in order to solve the problems of individual ICA method. These included the uncertainty of the output ICs order, the selection of the number of components, and the randomness of the initial value in the ICA algorithm, which could identify the reliable and stable ICs and obtain the intrinsic brain functional networks. Spatial ICA was performed on the preprocessed fMRI images for each participant. The obtained ICs were maps that were maximally spatially independent for each subject and represented the brain functional subnetworks. The mixing matrix represented time courses of the ICs, which represented the changes of the brain functional networks over time.

The number of ICs needs to be specified before ICA is performed. One cannot, however, know *a priori* the single number of components to generate with ICA that is "optimal" for the identification of reproducible components across individuals. Therefore, the principles of information criteria were applied to determine the number of sets of ICs in this study. We combined minimum description length (Calhoun et al., 2001) and Akaike's information criterion (Wang et al., 2011a) to estimate the interval (lower and upper bounds) and step size of the numbers of ICs. Additionally, the initial values of the ICA algorithm are random, meaning that the objective function in the ICA algorithm may fall into a different local extremum. As a result, the inconsistent ICs may be produced when the same ICA algorithm is performed on the same subject with the same number of components. Accordingly, in this study, the spatial ICA algorithm was run several times with the estimated numbers of ICs on each individual subject. Then the correspondence of ICs between different subjects with a set of numbers of ICs was established by the hierarchical partner matching method, which we proposed and published previously (Wang et al., 2011a; Qiao et al., 2015,

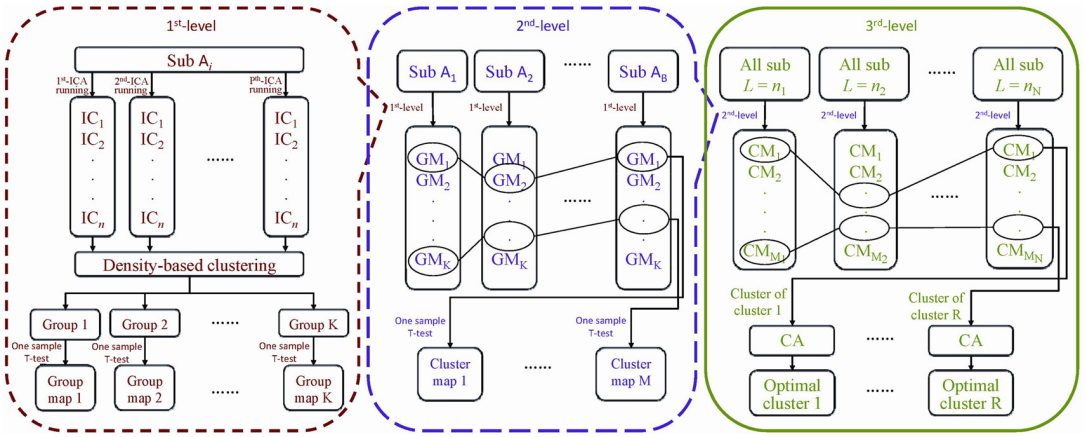

FIGURE 1 | The flowchart of the three-level hierarchical partner matching independent component analysis (3LHPM-ICA) algorithm.

2017). In detail, the proposed 3LHPM-ICA approach consists of three levels as follows and its framework is shown in **Figure 1**.

In the first level, in order to address the problem of the randomness of the initial values in the ICA algorithm, we inputted the fMRI data of each subject and performed spatial ICA by P multiplied with the single number of ICs. Then the ICs of the subject (denoted as subject A_j) were clustered by the density-based clustering algorithm which had high efficiency and low complexity (Rodriguez and Laio, 2014). Specifically, each IC was considered as one point in the high dimensional space. The local density of the point and its distance from points of higher density were computed for each data point. Here, the Pearson correlation coefficient was applied to measure the distance between two points. Then, the local density and distance of all points were sorted in descending order. The first K points were identified as center points. After that, the distances from all other points to the center points were calculated for group assignment. Finally, a group map (GM) was generated by running one-sample t-tests on each group of ICs.

In the second level, in order to solve the problem of the correspondence of ICs across different individuals, the GMs of all the subjects $\{A_1, A_2, \ldots, A_B\}$ that generated with the same single number of ICs were matched by the partner matching method, which we proposed and published previously (Wang and Peterson, 2008). The Tanimoto distance was used to measure the similarity between GMs. Given a GM_i of subject A_1, the indices of spatial similarity between GM_i and all the GMs of subject A_2 were calculated. The GM_j of subject A_2 was selected, which had the maximum similarity index with GM_i of subject A_1 among all the GMs of subject A_2. After that, the similarity indices between GM_j of subject A_2 and all the GMs of subject A_1 were calculated. The GM_k of subject A_1 was selected which had the maximum similarity index with GM_j of subject A_2 among all the GMs of subject A_1. If $k = i$, then the matching was bidirectional, and we considered GM_i of subject A_1 and GM_j of subject A_2 to be partner matched. This procedure was repeated to find all pairs of GMs that are bidirectionally matched between subject A_1 and A_2. Similarly, the partner matching method was performed to identify matching GMs across all the subjects. A collection of GMs that match across subjects was termed as a *cluster*. Finally, a cluster map (CM) was generated by running one-sample t-tests on each cluster of GMs, which represented a spatial pattern that tends to be present across subjects.

In the third level, in order to figure out the correspondence of ICs across different numbers, the CMs of all the subjects that generated with the estimated multiple numbers of ICs $L = \{n_1, n_2, \ldots, n_N\}$ were clustered by the partner matching method, identifying corresponding CMs across the different sets that were obtained with different numbers of ICs. For each cluster of CMs, the cluster with the highest Cronbach's Alpha was selected as the optimal cluster. The CMs were derived from GMs and GMs were derived from ICs, thus the most reliable and stable ICs could be obtained by backward tracing from optimal clusters.

Effective Connectivity Analysis Based on Granger Causality

GC has been widely applied to assess brain effective connectivity in fMRI data analysis. Compared with the structural equation model and dynamic causal model, GC analysis is very consistent with the actual situation because it considers time and does not require any prior knowledge (Goebel et al., 2003; Cohen Kadosh et al., 2016). In this study, we computed the GC index (GCI) to assess the causal influence between the ICs that were identified by the 3LHPM-ICA method.

Let $X(t)$ denote the zero-mean vector time course of an ICs within region X, and $Y(t)$ denote the zero-mean vector time course of another IC within region Y. Then $X(t)$ can be estimated by applying an autoregressive (AR) model of order P as follows:

$$X(t) = \sum_{i=1}^{P} \alpha_i X(t-i) + \varepsilon_X \quad (1)$$

where α_i are coefficients of the AR model and ε_X is the zero-mean residual. The $Y(t)$ is then added into the above AR model and

$X(t)$ can be estimated by

$$X(t) = \sum_{i=1}^{P} \alpha_i X(t-i) + \sum_{j=1}^{P} \beta_j Y(t-j) + \varepsilon_{XY} \quad (2)$$

where β_j are coefficients of the AR model and ε_{XY} is the new zero-mean residual. To assess whether the addition of $Y(t)$ improves the prediction compared with the use of $X(t)$ alone, the GCI from Y to X can be calculated by

$$GCI_{Y\to X} = 1 - \frac{var(\varepsilon_{XY})}{var(\varepsilon_X)} \quad (3)$$

where $var(\varepsilon_{XY})$ and (ε_X) are the variance of the estimation errors or residuals ε_{XY} and ε_X, respectively. If $GCI_{(Y\to X)}$ is greater than zero, the addition of the previous values of $Y(t)$ into the right side of Equation (1) significantly improves the prediction of the current values of $X(t)$ and we can deem that $Y(t)$ Granger caused $X(t)$, that is, region Y has a causal influence and directional interaction to region X.

In this way, a GCI matrix was obtained by repeating the above procedure to all ICs for each subject. In the GCI effective connectivity matrix, rows and columns of the matrix represented different ICs. Each cell of the matrix represented a distinct connection between two ICs corresponding to specific row and column. The diagonal value of the matrix was *NaN* because there was no meaningful directional interaction from one IC to the same one. The GCI matrices of all subjects were computed, which would be applied as an effective feature in the following classifier.

Feature Fusion and Classification

The deep learning classification framework in this study consists of four steps: multivariate analysis, feature extraction, feature fusion and directed acyclic graph (DAG) network, as shown in **Figure 2**. The details can be stated as follows. First, reproducible ICs were obtained by performing 3LHPM-ICA on training resting state fMRI data. Then the GCIs were computed to infer directional interaction between these brain regions by extracting the time course of each IC within each pattern. Second, the z-score maps of the reliable ICs were then entered into a two-sample t-test model implemented in the SPM12 factorial module to detect group difference of the FC between AD and NC. The ROIs with significant differences ($p < 0.05$, uncorrected) between the two groups of the training set were extracted as FC features for the pattern recognition analyses. In addition, GC matrices computed by the time course of significant ICs were selected as effective connectivity features. Third, functional and effective connectivity features were fused by replacing the diagonal values *NaN* in the GC matrices as IC features. In this way, a matrix feature was obtained for each subject. Finally, the two-dimensional characteristic matrices of the training data were inputted into a deep learning classifier model. Given test fMRI data, the same steps were conducted and a feature matrix was entered into the pretrained network for the prediction of AD/NC. A leave-one-out cross-validation (LOOCV) strategy was applied to evaluate the performance of the classifier.

A DAG network is a deep learning method which has its layers arranged as a DAG and a more complex architecture where layers can have inputs from, or outputs to, multiple layers. In this study, we implemented the DAG network for deep learning with the neural network toolbox in MATLAB R2018a, as shown in **Figure 3**, which consisted of a main branch with layers connected sequentially and a shortcut connection that enabled the parameter gradients to flow more easily from the output layer to the earlier layers of the network. The main branch contained an image input layer, three convolutional layers, three batch normalization layers, three rectified linear unit (ReLU) layers, an average pooling layer, a

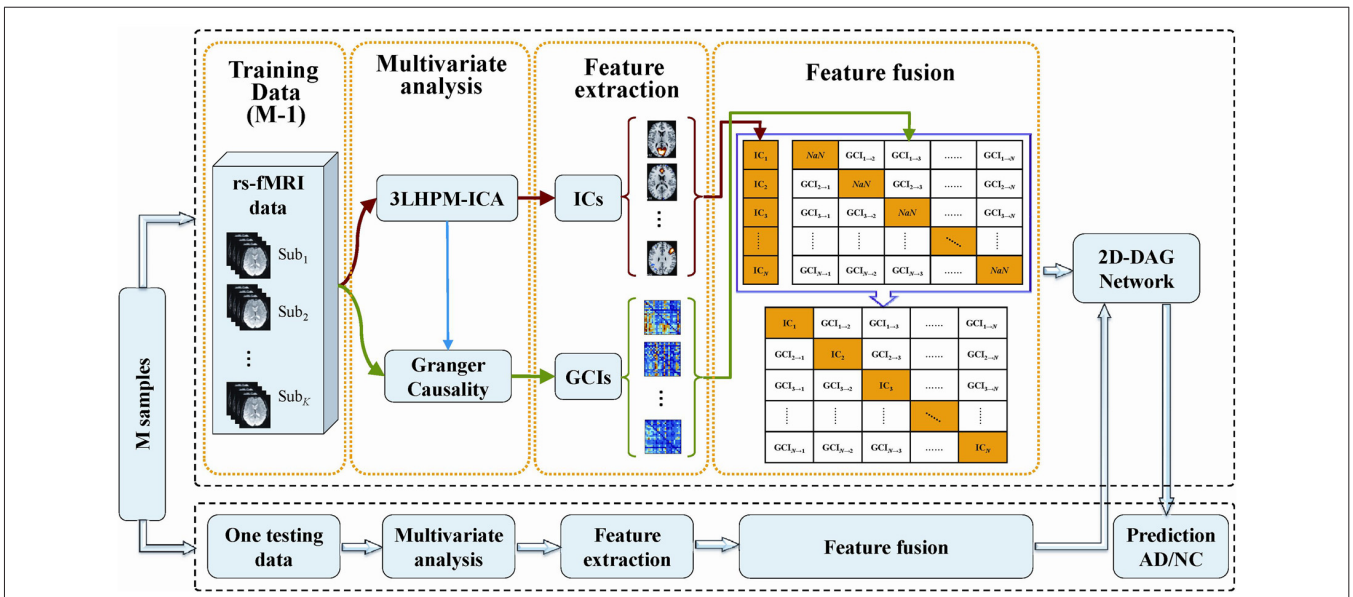

FIGURE 2 | The framework of the proposed deep learning classification algorithm based on 3LHPM-ICA and Granger causality (GC).

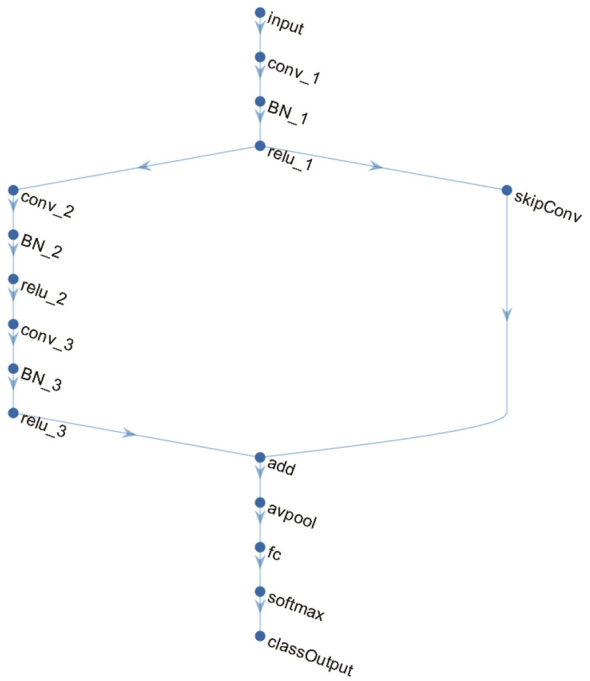

FIGURE 3 | The architecture of the directed acyclic graph (DAG) network.

fully connected layer, a softmax layer and classification layer. The shortcut connection contained a single one-by-one convolutional layer that had an added benefit of not adding any extra parameters or computational complexity. Batch normalization layers between convolutional layers and ReLU layers normalized the activations and gradients propagating through a network, resulting in speeding up network training and reducing the sensitivity to network initialization. The average pooling layer was applied as a down-sampling operation that reduced the spatial size of the feature map and removed redundant spatial information.

RESULTS

ICA-Based Functional Connectivity

We performed the 3LHPM-ICA method on the training fMRI data. The numbers of components were set to be 20 to 130, with increments of 10 which were determined by information criteria. In the first level, we performed 10 times ICA with the single number of ICs on the fMRI data of each subject. The first K points were identified as center points in the density-based clustering algorithm. The K was set to be n plus 10 experimentally, where n is the number of ICs. In the second level, we performed the partner matching method on the training subjects with the same single number of ICs. The numbers of the CMs were 29, 36, 46, 55, 62, 69, 77, 86, 95, 102, 113 and 122, while the numbers of the ICs were 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120 and 130, respectively. In the third level, 27 cluster of clusters were obtained after performing the partner matching method. Three artifactual cluster of clusters were excluded. Finally, 24 clusters of ICs that were significantly reproducible in their spatial patterns across individuals were identified. The general linear model in SPM was utilized to perform a one-sample t-test on each of the clusters to generate IC maps that represented FC features. After that, the reproducible ICs of AD and NC were compared in a second-level random effects analysis, covarying with age and sex. Compared with NC, FC in AD was significantly decreased in various cortical and subcortical areas related to memory, emotion and cognition, including the middle frontal gyrus (MFG), superior medial gyrus (SMG), middle orbital gyrus (MOG), inferior frontal gyrus (IFG), supplementary motor area (SMA), medial frontal gyrus (MedFG), hippocampus, insula, putamen, anterior cingulate cortex (ACC), posterior cingulate cortex (PCC), superior parietal lobule (SPL), superior temporal gyrus (STG), and middle temporal gyrus (MTG; **Figure 4**, **Table 1**).

GC Based Effective Connectivity

The effective connectivity was measured by computing the GC of time courses of 24 ICs identified by 3LHPM-ICA. The 24 × 24 GCI matrix was obtained for each subject. The diagonal of the GCI matrix was set to be *NaN* because there is no meaning for the GC from brain area X to itself. Finally, the functional and effective connectivity features were fused by replacing the diagonal values of the GCI matrix with IC values in the z-score IC maps.

Classification

We applied the DAG network for deep learning to classify and predict the AD/NC. The image size at the input layer in **Figure 3** was 24 × 24 × 1. The filter size in the convolutional layer "*conv_1*" was 5 × 5. The number of filters was 16, which represented the number of neurons that connect to the same region of the input. The filter size of "*conv_2*" and "*conv_3*" were 3 × 3 with 32 filters. The window size in the average pooling layer "*avpool*" was 3 × 3 with stride (or step size) 2 × 2. The filter size in the convolutional layer of the shortcut connection "*skipConv*" was 1 × 1 with 32 filters. The training lasted for 20 epochs. The batch size was 20. The iteration per epoch was three and the total iteration was 60. The initial learning rate was set to be 0.01. The learning rate was multiplied by a factor every time a certain number of epochs had passed. The multiplicative factor was 0.1 and the number of epochs between multiplications was 10. The output was a 1 × 2 vector containing the probabilities of the test data belonging to AD or NC.

In every fold of LOOCV, the number of the training data was 67 and the last one was used as testing data. In the training stage, we performed 3LHPM-ICA and GC on the 67 training data. The extracted features were then entered into the classifier model. In the testing stage, the ICA was performed on the testing data. Then the most similar ICs of the testing data were selected by computing the Euclidean distance between the ICs of the testing data and the reproducible ICs from the training data. Finally, the ROIs of the selected ICs and GCIs were entered into the classifier for the prediction of AD/NC. For each subject, the 24 by 24 feature matrix was entered into the

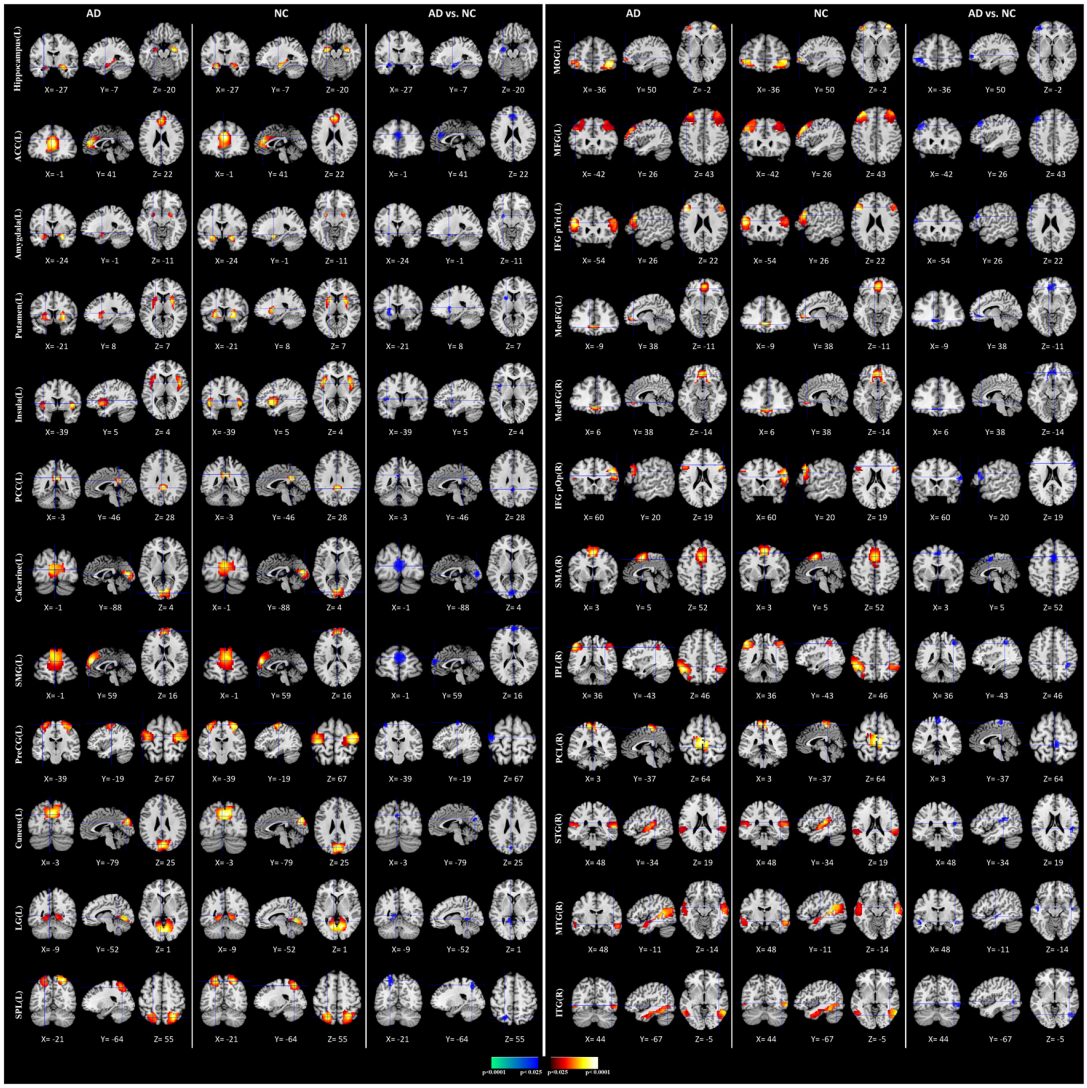

FIGURE 4 | Comparisons of functional connectivity (FC) between Alzheimer's disease (AD) and normal controls (NCs). The first and fourth columns of three display the random-effect group connectivity maps detected from the AD. Within each column of three, the first column is a coronal view, the second is a sagittal view, and the third is an axial view. The second and fifth columns of three display the group connectivity maps detected from the NCs. Each row displays one group connectivity map generated by applying a one-sample *t*-test to the clusters of ICs. Any two group connectivity maps within the same row across the first three and second three columns (as well as the fourth three and fifth three columns) are significantly similar to one another in their spatial configurations. The third and sixth columns of three display t-contrast maps comparing the group connectivity maps from the AD and control participants. MFG, middle frontal gyrus; MedFG, medial frontal gyrus; SMG, superior medial gyrus; MOG, middle orbital gyrus; IFG pOp, inferior frontal gyrus (p. Opercularis); IFG pTri, inferior frontal gyrus (p. Triangularis); SMA, supplementary motor area; ACC, anterior cingulate cortex; PCC, posterior cingulate cortex; SPL, superior parietal lobule, IPL, inferior parietal lobule; PCL, paracentral lobule; STG, superior temporal gyrus; MTG, middle temporal gyrus; ITG, inferior temporal gyrus; PreCG, precentral gyrus; LG, lingual gyrus.

deep learning network. With LOOCV strategy, a classification accuracy of 95.59% with a sensitivity of 97.06% and a specificity of 94.12% was achieved. For comparison the classifiers, including LeNet5 (LeCun et al., 1998), the kernel support vector machine (SVM), the maximum uncertainty linear discriminant analysis (MDLA; Dai et al., 2012) and autoencoder (AE), were also performed. The deep neural network with stacked AEs consisted of five layers: an input layer, two hidden layers, a softmax layer and a classification layer. First, we trained the hidden layers individually in an unsupervised fashion using AEs. Then we trained a softmax layer and joined the layers together to form a stacked network. Finally, a supervised fine-tuning stage was applied to improve the classification performance by performing backpropagation on the whole multilayer network. The numbers of nodes were set to be 100 and 50 in the first and second hidden layers, respectively. A Gaussian kernel with a width of 0.5 was used in SVM. Several types of features, including the AAL atlas-based features, GC features and combined ICA and GC features with different classifiers were also implemented. The AAL atlas-based features were 90 × 90 matrices obtained

TABLE 1 | Location and comparisons of independent component (IC) maps between Alzheimer's disease (AD) and normal control (NC).

Brain areas	Location		Peak location			textitT
	Side	BA	*x*	textity	textitz	statistic
AD vs. NC (negative)						
Middle frontal gyrus	L	8	−42	26	43	−4.06
Superior medial gyrus	L	10	−1	59	16	−4.05
Calcarine gyrus	L	17	−1	−88	4	−4.00
Middle orbital gyrus	L	10	−36	50	−2	−3.67
Inferior frontal gyrus (p. Opercularis)	R	44	60	20	19	−3.48
Inferior frontal gyrus (p. Triangularis)	L	45	−54	26	22	−3.28
Supplementary motor area	R	6	3	5	52	−3.03
Precentral gyrus	L	6	−39	−19	67	−3.08
Medial frontal gyrus	L	11	−9	38	−11	−3.97
	R	11	6	38	−14	−3.35
Insula	L	13	−39	5	4	−2.56
Anterior cingulate cortex	L	32	−1	41	22	−6.55
Posterior cingulate cortex	L	23	−3	−46	28	−2.94
Hippocampus	L	54	−27	−7	−20	−4.60
Amygdala	L	53	−24	−1	−11	−3.86
Putamen	L	49	−21	8	7	−2.92
Cuneus	L	18	−3	−79	25	−2.48
Lingual gyrus	L	18	−9	−52	1	−2.84
Superior parietal lobule	L	7	−21	−64	55	−2.63
Inferior parietal lobule	R	7	36	−43	46	−3.41
Paracentral lobule	R	4	3	−37	64	−3.11
Superior temporal gyrus	R	22	48	−34	19	−3.23
Middle temporal gyrus	R	22	48	−11	−14	−3.35
Inferior temporal gyrus	R	20	44	−67	−5	−2.35

All coordinates are in the Montreal Neurological Institute (MNI) ICBM 152 template.

by calculating the Pearson correlation coefficients between the brain regions, excluding the cerebellum, that were defined with AAL atlas. The upper triangular feature matrices were reshaped as feature vectors when SVM and MDLA were performed. The classification results are shown in **Table 2**. It can be seen that the classification performance of the DAG network combined with ICA and GC features is better than the values obtained with any single type of features or other types of classifiers.

The weights of the features were computed by the coefficients of the discrimination hyperplane, and the most discriminative features for classification are shown in **Figure 5**. The connections with the largest weights are the most informative. It can be seen that the IC activity in the MOG, IFG, MFG, ACC, insula, hippocampus, STG, and the effective connections from IFG to hippocampus, from ITG to precentral gyrus (PreCG), and from MFG to hippocampus made larger contributions to the classification.

TABLE 2 | Classification performance of different methods with leave-one-out cross validation (LOOCV).

Methods	Accuracy	Sensitivity	Specificity
AAL atlas based+SVM	77.94%	73.53%	82.35%
AAL atlas based+MDLA	75.0%	79.41%	70.59%
AAL atlas based+LeNet5	79.41%	76.47%	82.35%
AAL atlas based+AE	80.88%	76.47%	85.29%
AAL atlas based+DAG	82.35%	79.41%	85.29%
GC+SVM	83.82%	85.29%	82.35%
GC+MDLA	82.35%	88.24%	76.47%
GC+LeNet5	85.29%	82.35%	88.24%
GC+AE	88.24%	82.35%	94.12%
GC+DAG	88.24%	91.18%	85.29%
ICA+GC+SVM	91.18%	88.24%	94.12%
ICA+GC+MDLA	89.71%	97.06%	82.35%
ICA+GC+LeNet5	92.65%	94.12%	91.18%
ICA+GC+AE	94.12%	97.06%	91.18%
ICA+GC+DAG	95.59%	97.06%	94.12%

DISCUSSION

In the current work, we presented a 3LHPM-ICA approach which addressed the problems in spatial individual ICA algorithm such as the uncertainty of the number of components, the randomness of initial values, and the correspondence of ICs among multiple subjects. Then, we applied the 3LHPM-ICA method and GC on resting state fMRI data to investigate the reproducible and stable ICs across individuals. We then obtained the intrinsic brain functional and effective connectivity feature matrices. A deep learning framework was finally investigated to assess if these brain features can serve as biomarkers for AD.

We found significantly decreased intrinsic FC in AD patients compared to NC in several subcortical regions including the hippocampus, amygdala, insula and putamen. As one of the earliest and most widely investigated brain regions in AD, researchers have correlated alterations in hippocampal activity and connectivity as well as shrinkage with the presence of AD, which explains one of the early symptoms in the impairment of memory, especially the formation of new memories in AD patients (Wang L. et al., 2006; Allen et al., 2007; Mu and Gage, 2011; Smith et al., 2014). Amygdala atrophy in AD and its relation to global illness severity have also been reported (Scott et al., 1991; Barnes et al., 2006; Poulin et al., 2011), elucidating the aberrant motor behavior, anxiety and irritability of AD patients. Another positron emission tomographic study of AD reported the cholinergic deficit in the amygdala, supporting that the amygdala played an important role in the retention of affective conditioning and/or memory consolidation and cross-verified the role of the amygdala in the emotional and behavioral symptoms of AD (Shinotoh et al., 2003). The insula is a key region for cognition, emotion and sensory processes which has been demonstrated with gray matter loss (Guo et al., 2012), abnormal activities (Lin et al., 2017), and disrupted connections in AD (Xie et al., 2012; Liu et al., 2018). Furthermore, the reduced volumes of putamen, which was correlated with impaired global cognitive performance, might contribute to cognitive decline in AD (de Jong et al., 2008; Roh et al., 2011). Consistent with the previous studies, our findings of decreased brain connectivity in certain subcortical areas indicated that these alterations might be related to the memory, emotion, motor and cognition disorders present in AD patients.

The loss of neurons and synapses in the cerebral cortex of AD results in gross atrophy of the affected regions, including degeneration in the temporal gyrus, parietal lobe, and parts of the frontal cortex and cingulate gyrus. Neuropathological studies have shown that AD-related degeneration begins in the medial temporal lobe (Braak and Braak, 1995). The current finding of decreased FC in the temporal gyrus is in line with previous

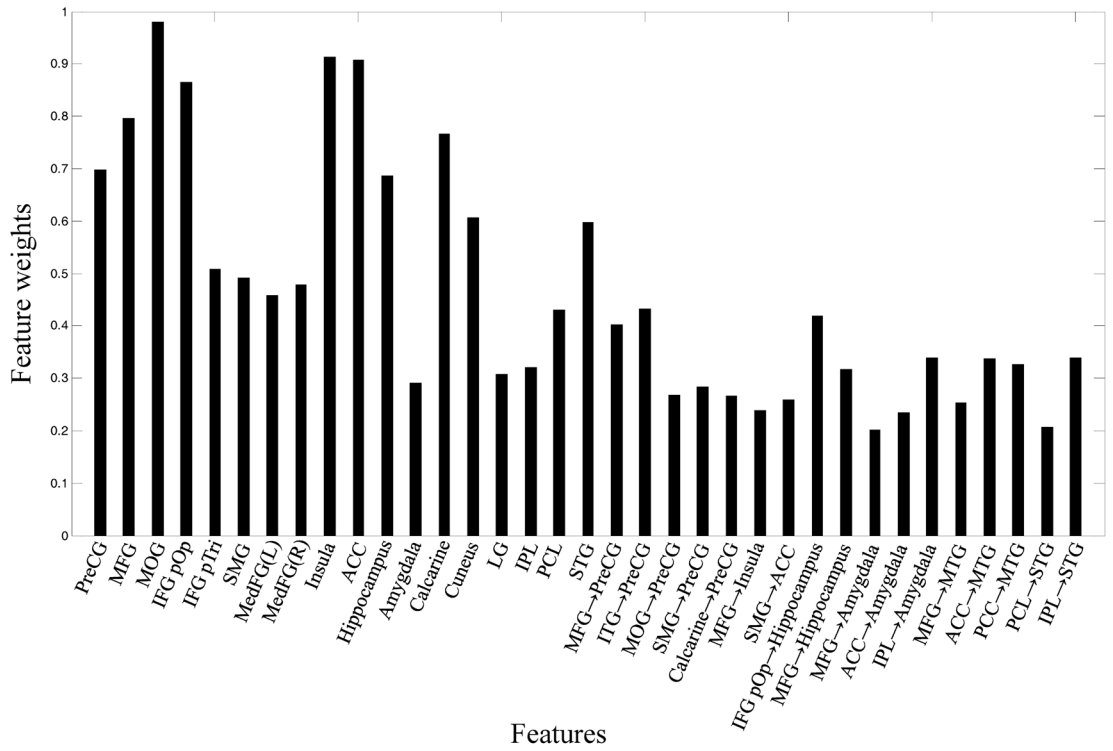

FIGURE 5 | Feature weights in the classification.

reports of temporal gyrus atrophy (Farrow et al., 2007; Frisoni et al., 2010; Ho et al., 2010) and FC anomalies (Toussaint et al., 2014), leading to the memory and learning deficits that are classically observed with early clinical manifestations of AD. Our results also revealed disrupted resting state functional connectivities in the DMN, which consists of the PCC, inferior parietal lobe (IPL) and prefrontal cortex (PFC). The cortical thinning (Dickerson and Sperling, 2009) and decreased intrinsic brain activity (He et al., 2007; Wang et al., 2011b) and connectivity (Greicius et al., 2004; Toussaint et al., 2014) of DMN have been demonstrated in many studies. Therefore, our findings provide further evidence that the aberration of DMN may result in the episodic memory, visual imagery and mentalizing disorders in AD. Moreover, as part of the frontostriatal circuit which is composed of the ACC, PFC and parts of the basal ganglia, the ACC is involved in effort-based decision making and executive functions (Stella et al., 2014; Theleritis et al., 2014; Le Heron et al., 2018). Disruption of the FC in ACC found in this study might play a pivotal role in apathy, such as behavioral activation, social motivation and emotional sensitivity disorders in AD patients. Therefore, the brain connectivity alterations of the identified cortical and subcortical regions in this study may be associated with the cognitive and functional impairment of AD and potentially served as clinical biomarkers of AD.

The two-dimensional features fused by the FC obtained by 3LHPM-ICA and effective connectivity derived from GC were then applied for classification in this study. Compared with the traditional feature arrangement and fusion method, which usually reshaped the two dimensional features into a vector or concatenated different types of features into a longer feature vector (Wang K. et al., 2006; Chen et al., 2011; Dai et al., 2012; Dyrba et al., 2015; de Vos et al., 2018), the two dimensional feature matrices and feature fusion method used in this study preserved the spatial structural characteristics of features and provided a more meaningful way to combine various types of features for classification. Moreover, the overfitting issue, which may be caused by high-dimensional feature space in the traditional methods, could be alleviated due to the two dimensions of features in this study.

Advanced deep learning techniques have been successfully applied for the diagnosis of AD based on PET and sMRI (Suk and Shen, 2013; Liu S. et al., 2015; Ortiz et al., 2016; Lu et al., 2018; Shi et al., 2018). A recent report constructed a customized AE architecture with resting-state correlation based FC to classify mild cognitive impairments from NCs (Ju et al., 2017). However, different parcellation schemes may generate different results. Therefore, compared with the correlation-based method, the data-driven method in this study avoided the problem whereby the brain parcellation methods may affect classification performance. The connectivity patterns of brain networks derived from ICA and GC were stable and not influenced by different parcellation atlases. Moreover, we

compared two kinds of deep learning algorithms with the same inputted features. One was LeNet5 with sequential connected layers and the other was the DAG network, which consisted of sequential connected layers and shortcut connections. Our results demonstrated that the DAG network has better performance than the sequential network, possibly because of the "skip" connections between layers with feed-forward computations.

Several limitations of the present study should be noted. First, the sample size in this study was not large and future work should be done on a larger training sample in order to improve the robustness and generalization of the classification model. Second, multimodal neuroimaging features such as sMRI and DTI should also be investigated in addition the resting state fMRI, which may lead to higher classification accuracy. Third, we used a binary classification for the prediction of AD/NC. However, multi-class classification should be considered for its clinical applications in the future because there are different stages of AD such as MCI, LMCI and EMCI. Fourth, it would be more comparable to compare the accuracy results with the same benchmark datasets. Therefore, future work will focus on the implementation of different models based on public datasets such as ADNI. Finally, a light deep architecture with two-dimensional input images was applied in this study. More complicated deep learning models should be implemented such as GoogLeNet, AlexNet, VGG, ResNet and 3D convolutional neural networks, which may be more appropriate for big data. Nevertheless, our results suggested that the functional and effective connectivity features extracted by 3LHPM-ICA and GC followed by deep learning classification represented the most powerful method of distinguishing AD from healthy data. Due to the flexibility of this technique, it has the potential to be extended to other psychiatric disorders in the future.

AUTHOR CONTRIBUTIONS

JQ, YL, and AL conceived and designed the experiments and performed the experiments. JQ, ZW and AL analyzed the data and contributed reagents, materials and analysis tools. JQ, CC, and AL wrote the article.

FUNDING

This work was supported by the National Natural Science Foundation of China (61603225), Natural Science Foundation of Shandong Province (ZR2016FQ04), China Postdoctoral Science Foundation (2016M602182), Key Research and Development Foundation of Shandong Province (2016GGX101009), Natural Science Foundation of Shandong Province (ZR2014FM012), Shandong Provincial Key Research and Development Plan (2017CXGC1504), and Natural Science Foundation for Distinguished Young Scholars of Shandong Province (JQ201718).

REFERENCES

Allen, G., Barnard, H., McColl, R., Hester, A. L., Fields, J. A., Weiner, M. F., et al. (2007). Reduced hippocampal functional connectivity in Alzheimer disease. *Arch. Neurol.* 64, 1482–1487. doi: 10.1001/archneur.64.10.1482

Alzheimer's Association. (2011). 2011 Alzheimer's disease facts and figures. *Alzheimers Dement.* 7, 208–244. doi: 10.1016/j.jalz.2011.02.004

Babiloni, C., Triggiani, A. I., Lizio, R., Cordone, S., Tattoli, G., Bevilacqua, V., et al. (2016). Classification of single normal and Alzheimer's disease individuals from cortical sources of resting state EEG rhythms. *Front. Neurosci.* 10:47. doi: 10.3389/fnins.2016.00047

Barnes, J., Whitwell, J. L., Frost, C., Josephs, K. A., Rossor, M., and Fox, N. C. (2006). Measurements of the amygdala and hippocampus in pathologically confirmed Alzheimer disease and frontotemporal lobar degeneration. *Arch. Neurol.* 63, 1434–1439. doi: 10.1001/archneur.63.10.1434

Binnewijzend, M. A., Schoonheim, M. M., Sanz-Arigita, E., Wink, A. M., van der Flier, W. M., Tolboom, N., et al. (2012). Resting-state fMRI changes in Alzheimer's disease and mild cognitive impairment. *Neurobiol. Aging* 33, 2018–2028. doi: 10.1016/j.neurobiolaging.2011.07.003

Braak, H., and Braak, E. (1995). Staging of Alzheimer's disease-related neurofibrillary changes. *Neurobiol. Aging* 16, 271–278; discussion 278–284. doi: 10.1016/0197-4580(95)00021-6

Burns, A., and Iliffe, S. (2009). Alzheimer's disease. *BMJ* 338:b158. doi: 10.1136/bmj.b158

Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Hum. Brain Mapp.* 14, 140–151. doi: 10.1002/hbm.1048

Calhoun, V. D., Liu, J., and Adali, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage* 45, S163–S172. doi: 10.1016/j.neuroimage.2008.10.057

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., and Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *Neuroimage* 112, 232–243. doi: 10.1016/j.neuroimage.2015.02.037

Chen, G., Ward, B. D., Xie, C., Li, W., Wu, Z., Jones, J. L., et al. (2011). Classification of Alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging. *Radiology* 259, 213–221. doi: 10.1148/radiol.10100734

Chupin, M., Gerardin, E., Cuingnet, R., Boutet, C., Lemieux, L., Lehéricy, S., et al. (2009). Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus* 19, 579–587. doi: 10.1002/hipo.20626

Cohen Kadosh, K., Luo, Q., de Burca, C., Sokunbi, M. O., Feng, J., Linden, D. E. J., et al. (2016). Using real-time fMRI to influence effective connectivity in the developing emotion regulation network. *Neuroimage* 125, 616–626. doi: 10.1016/j.neuroimage.2015.09.070

Colliot, O., Chetelat, G., Chupin, M., Desgranges, B., Magnin, B., Benali, H., et al. (2008). Discrimination between Alzheimer disease, mild cognitive impairment, and normal aging by using automated segmentation of the hippocampus. *Radiology* 248, 194–201. doi: 10.1148/radiol.2481070876

Craddock, R. C., James, G. A., Holtzheimer, P. E. III., Hu, X. P., and Mayberg, H. S. (2012). A whole brain fMRI atlas generated via spatially constrained spectral clustering. *Hum. Brain Mapp.* 33, 1914–1928. doi: 10.1002/hbm.21333

Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., et al. (2012). Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3). *Neuroimage* 59, 2187–2195. doi: 10.1016/j.neuroimage.2011.10.003

de Jong, L. W., van der Hiele, K., Veer, I. M., Houwing, J. J., Westendorp, R. G., Bollen, E. L., et al. (2008). Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. *Brain* 131, 3277–3285. doi: 10.1093/brain/awn278

de Vos, F., Koini, M., Schouten, T. M., Seiler, S., van der Grond, J., Lechner, A., et al. (2018). A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *Neuroimage* 167, 62–72. doi: 10.1016/j.neuroimage.2017.11.025

Desikan, R. S., Cabral, H. J., Hess, C. P., Dillon, W. P., Glastonbury, C. M., Weiner, M. W., et al. (2009). Automated MRI measures identify individuals

with mild cognitive impairment and Alzheimer's disease. *Brain* 132, 2048–2057. doi: 10.1093/brain/awp123

Dickerson, B. C., Bakkour, A., Salat, D. H., Feczko, E., Pacheco, J., Greve, D. N., et al. (2009). The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals. *Cereb. Cortex* 19, 497–510. doi: 10.1093/cercor/bhn113

Dickerson, B. C., and Sperling, R. A. (2009). Large-scale functional brain network abnormalities in Alzheimer's disease: insights from functional neuroimaging. *Behav. Neurol.* 21, 63–75. doi: 10.3233/BEN-2009-0227

Diehl, J., Grimmer, T., Drzezga, A., Riemenschneider, M., Förstl, H., and Kurz, A. (2004). Cerebral metabolic patterns at early stages of frontotemporal dementia and semantic dementia. A PET study. *Neurobiol. Aging* 25, 1051–1056. doi: 10.1016/j.neurobiolaging.2003.10.007

Dubois, B., Feldman, H. H., Jacova, C., Hampel, H., Molinuevo, J. L., Blennow, K., et al. (2014). Advancing research diagnostic criteria for Alzheimer's disease: the IWG-2 criteria. *Lancet Neurol.* 13, 614–629. doi: 10.1016/S1474-4422(14)70090-0

Dyrba, M., Ewers, M., Wegrzyn, M., Kilimann, I., Plant, C., Oswald, A., et al. (2013). Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data. *PLoS One* 8:e64925. doi: 10.1371/journal.pone.0064925

Dyrba, M., Grothe, M., Kirste, T., and Teipel, S. J. (2015). Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum. Brain Mapp.* 36, 2118–2131. doi: 10.1002/hbm.22759

Ebadi, A., Dalboni da Rocha, J. L., Nagaraju, D. B., Tovar-Moll, F., Bramati, I., Coutinho, G., et al. (2017). Ensemble classification of Alzheimer's disease and mild cognitive impairment based on complex graph measures from diffusion tensor images. *Front. Neurosci.* 11:56. doi: 10.3389/fnins.2017.00056

Fan, Y., Batmanghelich, N., Clark, C. M., and Davatzikos, C. (2008). Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage* 39, 1731–1743. doi: 10.1016/j.neuroimage.2007.10.031

Farrow, T. F., Thiyagesh, S. N., Wilkinson, I. D., Parks, R. W., Ingram, L., and Woodruff, P. W. (2007). Fronto-temporal-lobe atrophy in early-stage Alzheimer's disease identified using an improved detection methodology. *Psychiatry Res.* 155, 11–19. doi: 10.1016/j.pscychresns.2006.12.013

Frisoni, G. B., Fox, N. C., Jack, C. R. Jr., Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77. doi: 10.1038/nrneurol.2009.215

Goebel, R., Roebroeck, A., Kim, D. S., and Formisano, E. (2003). Investigating directed cortical interactions in time-resolved fMRI data using vector autoregressive modeling and Granger causality mapping. *Magn. Reson. Imaging* 21, 1251–1261. doi: 10.1016/j.mri.2003.08.026

Greicius, M. D., Srivastava, G., Reiss, A. L., and Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc. Natl. Acad. Sci. U S A* 101, 4637–4642. doi: 10.1073/pnas.0308627101

Guo, X., Han, Y., Chen, K., Wang, Y., and Yao, L. (2012). Mapping joint grey and white matter reductions in Alzheimer's disease using joint independent component analysis. *Neurosci. Lett.* 531, 136–141. doi: 10.1016/j.neulet.2012.10.038

He, Y., Wang, L., Zang, Y., Tian, L., Zhang, X., Li, K., et al. (2007). Regional coherence changes in the early stages of Alzheimer's disease: a combined structural and resting-state functional MRI study. *Neuroimage* 35, 488–500. doi: 10.1016/j.neuroimage.2006.11.042

Ho, A. J., Hua, X., Lee, S., Leow, A. D., Yanovsky, I., Gutman, B., et al. (2010). Comparing 3 T and 1.5 T MRI for tracking Alzheimer's disease progression with tensor-based morphometry. *Hum. Brain Mapp.* 31, 499–514. doi: 10.1002/hbm.20882

Hua, X., Leow, A. D., Lee, S., Klunder, A. D., Toga, A. W., Lepore, N., et al. (2008). 3D characterization of brain atrophy in Alzheimer's disease and mild cognitive impairment using tensor-based morphometry. *Neuroimage* 41, 19–34. doi: 10.1016/j.neuroimage.2008.02.010

Jack, C. R. Jr., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., et al. (2008). The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J. Magn. Reson. Imaging* 27, 685–691. doi: 10.1002/jmri.21049

Ju, R., Hu, C., Zhou, P., and Li, Q. (2017). Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/tcbb.2017.2776910 [Epub ahead of print].

Khazaee, A., Ebrahimzadeh, A., and Babajani-Feremi, A. (2015). Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. *Clin. Neurophysiol.* 126, 2132–2141. doi: 10.1016/j.clinph.2015.02.060

Khazaee, A., Ebrahimzadeh, A., and Babajani-Feremi, A. (2016). Application of advanced machine learning methods on resting-state fMRI network for identification of mild cognitive impairment and Alzheimer's disease. *Brain Imaging Behav.* 10, 799–817. doi: 10.1007/s11682-015-9448-7

Khazaee, A., Ebrahimzadeh, A., and Babajani-Feremi, A. (2017). Classification of patients with MCI and AD from healthy controls using directed graph measures of resting-state fMRI. *Behav. Brain Res.* 322, 339–350. doi: 10.1016/j.bbr.2016.06.043

Le Heron, C., Apps, M. A. J., and Husain, M. (2018). The anatomy of apathy: a neurocognitive framework for amotivated behaviour. *Neuropsychologia* 118, 54–67. doi: 10.1016/j.neuropsychologia.2017.07.003

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lehmann, C., Koenig, T., Jelic, V., Prichep, L., John, R. E., Wahlund, L. O., et al. (2007). Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *J. Neurosci. Methods* 161, 342–350. doi: 10.1016/j.jneumeth.2006.10.023

Lerch, J. P., Pruessner, J., Zijdenbos, A. P., Collins, D. L., Teipel, S. J., Hampel, H., et al. (2008). Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls. *Neurobiol. Aging* 29, 23–30. doi: 10.1016/j.neurobiolaging.2006.09.013

Li, S. J., Li, Z., Wu, G., Zhang, M. J., Franczak, M., and Antuono, P. G. (2002). Alzheimer disease: evaluation of a functional MR imaging index as a marker. *Radiology* 225, 253–259. doi: 10.1148/radiol.2251011301

Lin, F., Ren, P., Lo, R. Y., Chapman, B. P., Jacobs, A., Baran, T. M., et al. (2017). Insula and inferior frontal gyrus' activities protect memory performance against Alzheimer's disease pathology in old age. *J. Alzheimers Dis.* 55, 669–678. doi: 10.3233/jad-160715

Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Feng, D., et al. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Trans. Biomed. Eng.* 62, 1132–1140. doi: 10.1109/tbme.2014.2372011

Liu, X., Chen, X., Zheng, W., Xia, M., Han, Y., Song, H., et al. (2018). Altered functional connectivity of insular subregions in Alzheimer's disease. *Front. Aging Neurosci.* 10:107. doi: 10.3389/fnagi.2018.00107

Liu, M., Zhang, D., and Shen, D. (2015). View-centralized multi-atlas classification for Alzheimer's disease diagnosis. *Hum. Brain Mapp.* 36, 1847–1865. doi: 10.1002/hbm.22741

Liu, M., Zhang, D., and Shen, D. (2016). Relationship induced multi-template learning for diagnosis of Alzheimer's disease and mild cognitive impairment. *IEEE Trans. Med. Imaging* 35, 1463–1474. doi: 10.1109/TMI.2016.2515021

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., and Beg, M. F. (2018). Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Sci. Rep.* 8:5697. doi: 10.1038/s41598-018-22871-z

Magnin, B., Mesrob, L., Kinkingnéhun, S., Pélégrini-Issac, M., Colliot, O., Sarazin, M., et al. (2009). Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. *Neuroradiology* 51, 73–83. doi: 10.1007/s00234-008-0463-x

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of department of health and human services task force on Alzheimer's disease. *Neurology* 34, 939–944. doi: 10.1212/wnl.34.7.939

Mirzaei, G., Adeli, A., and Adeli, H. (2016). Imaging and machine learning techniques for diagnosis of Alzheimer's disease. *Rev. Neurosci.* 27, 857–870. doi: 10.1515/revneuro-2016-0029

Morris, J. C. (1993). The clinical dementia rating (CDR): current version and scoring rules. *Neurology* 43, 2412–2414. doi: 10.1212/wnl.43.11.2412-a

Mu, Y., and Gage, F. H. (2011). Adult hippocampal neurogenesis and its role in Alzheimer's disease. *Mol. Neurodegener.* 6:85. doi: 10.1186/1750-1326-6-85

Ortiz, A., Munilla, J., Górriz, J. M., and Ramírez, J. (2016). Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. *Int. J. Neural Syst.* 26:1650025. doi: 10.1142/s0129065716500258

Poulin, S. P., Dautoff, R., Morris, J. C., Barrett, L. F., and Dickerson, B. C. (2011). Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Res.* 194, 7–13. doi: 10.1016/j.pscychresns.2011.06.014

Qiao, J., Wang, Z., Zhao, G., Huo, Y., Herder, C. L., Sikora, C. O., et al. (2017). Functional neural circuits that underlie developmental stuttering. *PLoS One* 12:e0179255. doi: 10.1371/journal.pone.0179255

Qiao, J., Weng, S., Wang, P., Long, J., and Wang, Z. (2015). Normalization of intrinsic neural circuits governing Tourette's syndrome using cranial electrotherapy stimulation. *IEEE Trans. Biomed. Eng.* 62, 1272–1280. doi: 10.1109/tbme.2014.2385151

Querfurth, H. W., and LaFerla, F. M. (2010). Alzheimer's disease. *N. Engl. J. Med.* 362, 329–344. doi: 10.1056/NEJMra0909142

Rodriguez, A., and Laio, A. (2014). Machine learning. Clustering by fast search and find of density peaks. *Science* 344, 1492–1496. doi: 10.1126/science.1242072

Roh, J. H., Qiu, A., Seo, S. W., Soon, H. W., Kim, J. H., Kim, G. H., et al. (2011). Volume reduction in subcortical regions according to severity of Alzheimer's disease. *J. Neurol.* 258, 1013–1020. doi: 10.1007/s00415-010-5872-1

Scott, S. A., DeKosky, S. T., and Scheff, S. W. (1991). Volumetric atrophy of the amygdala in Alzheimer's disease: quantitative serial reconstruction. *Neurology* 41, 351–356. doi: 10.1212/wnl.41.3.351

Shi, J., Zheng, X., Li, Y., Zhang, Q., and Ying, S. (2018). Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J. Biomed. Health Inform.* 22, 173–183. doi: 10.1109/jbhi.2017.2655720

Shinotoh, H., Fukushi, K., Nagatsuka, S., Tanaka, N., Aotsuka, A., Ota, T., et al. (2003). The amygdala and Alzheimer's disease: positron emission tomographic study of the cholinergic system. *Ann. N Y Acad. Sci.* 985, 411–419. doi: 10.1111/j.1749-6632.2003.tb07097.x

Smith, J. C., Nielson, K. A., Woodard, J. L., Seidenberg, M., Durgerian, S., Hazlett, K. E., et al. (2014). Physical activity reduces hippocampal atrophy in elders at genetic risk for Alzheimer's disease. *Front. Aging Neurosci.* 6:61. doi: 10.3389/fnagi.2014.00061

Stella, F., Radanovic, M., Aprahamian, I., Canineu, P. R., de Andrade, L. P., and Forlenza, O. V. (2014). Neurobiological correlates of apathy in Alzheimer's disease and mild cognitive impairment: a critical review. *J. Alzheimers Dis.* 39, 633–648. doi: 10.3233/jad-131385

Suk, H. I., and Shen, D. (2013). Deep learning-based feature representation for AD/MCI classification. *Med. Image Comput. Comput. Assist. Interv.* 16, 583–590. doi: 10.1007/978-3-642-40763-5_72

Theleritis, C., Politis, A., Siarkos, K., and Lyketsos, C. G. (2014). A review of neuroimaging findings of apathy in Alzheimer's disease. *Int. Psychogeriatr.* 26, 195–207. doi: 10.1017/s1041610213001725

Thompson, P. M., Hayashi, K. M., Sowell, E. R., Gogtay, N., Giedd, J. N., Rapoport, J. L., et al. (2004). Mapping cortical change in Alzheimer's disease, brain development, and schizophrenia. *Neuroimage* 23, S2–S18. doi: 10.1016/j.neuroimage.2004.07.071

Toussaint, P. J., Maiz, S., Coynel, D., Doyon, J., Messé, A., de Souza, L. C., et al. (2014). Characteristics of the default mode functional connectivity in normal ageing and Alzheimer's disease using resting state fMRI with a combined approach of entropy-based and graph theoretical measurements. *Neuroimage* 101, 778–786. doi: 10.1016/j.neuroimage.2014.08.003

Triggiani, A. I., Bevilacqua, V., Brunetti, A., Lizio, R., Tattoli, G., Cassano, F., et al. (2017). Classification of healthy subjects and Alzheimer's disease patients with dementia from cortical sources of resting state EEG rhythms: a study using artificial neural networks. *Front. Neurosci.* 10:604. doi: 10.3389/fnins.2016.00604

Tsao, S., Gajawelli, N., Zhou, J., Shi, J., Ye, J., Wang, Y., et al. (2017). Feature selective temporal prediction of Alzheimer's disease progression using hippocampus surface morphometry. *Brain Behav.* 7:e00733. doi: 10.1002/brb3.733

Wang, K., Jiang, T., Liang, M., Wang, L., Tian, L., Zhang, X., et al. (2006). Discriminative analysis of early Alzheimer's disease based on two intrinsically anti-correlated networks with resting-state fMRI. *Med. Image Comput. Comput. Assist. Interv.* 9, 340–347. doi: 10.1007/11866763_42

Wang, L., Zang, Y., He, Y., Liang, M., Zhang, X., Tian, L., et al. (2006). Changes in hippocampal connectivity in the early stages of Alzheimer's disease: evidence from resting state fMRI. *Neuroimage* 31, 496–504. doi: 10.1016/j.neuroimage.2005.12.033

Wang, Z., Maia, T. V., Marsh, R., Colibazzi, T., Gerber, A., and Peterson, B. S. (2011a). The neural circuits that generate tics in Tourette's syndrome. *Am. J. Psychiatry* 168, 1326–1337. doi: 10.1176/appi.ajp.2011.09111692

Wang, Z., Yan, C., Zhao, C., Qi, Z., Zhou, W., Lu, J., et al. (2011b). Spatial patterns of intrinsic brain activity in mild cognitive impairment and Alzheimer's disease: a resting-state functional MRI study. *Hum. Brain Mapp.* 32, 1720–1740. doi: 10.1002/hbm.21140

Wang, Z., and Peterson, B. S. (2008). Partner-matching for the automated identification of reproducible ICA components from fMRI datasets: algorithm and validation. *Hum. Brain Mapp.* 29, 875–893. doi: 10.1002/hbm.20434

Wang, J., Wang, L., Zang, Y., Yang, H., Tang, H., Gong, Q., et al. (2009). Parcellation-dependent small-world brain functional networks: a resting-state fMRI study. *Hum. Brain Mapp.* 30, 1511–1523. doi: 10.1002/hbm.20623

Xie, C., Bai, F., Yu, H., Shi, Y., Yuan, Y., Chen, G., et al. (2012). Abnormal insula functional network is associated with episodic memory decline in amnestic mild cognitive impairment. *Neuroimage* 63, 320–327. doi: 10.1016/j.neuroimage.2012.06.062

Zalesky, A., Fornito, A., Harding, I. H., Cocchi, L., Yücel, M., Pantelis, C., et al. (2010). Whole-brain anatomical networks: does the choice of nodes matter? *Neuroimage* 50, 970–983. doi: 10.1016/j.neuroimage.2009.12.027

Zhang, D., and Raichle, M. E. (2010). Disease and the brain's dark energy. *Nat. Rev. Neurol.* 6, 15–28. doi: 10.1038/nrneurol.2009.198

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 11 February 2019
doi: 10.3389/fnagi.2019.00008

Evaluation of Functional Decline in Alzheimer's Dementia Using 3D Deep Learning and Group ICA for rs-fMRI Measurements

Muhammad Naveed Iqbal Qureshi[1,2,3,4†], *Seungjun Ryu*[1†], *Joonyoung Song*[1], *Kun Ho Lee*[5,6*] *and Boreom Lee*[1*]

[1]*Department of Biomedical Science and Engineering (BMSE), Institute of Integrated Technology (IIT), Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea,* [2]*Translational Neuroimaging Laboratory, The McGill University Research Centre for Studies in Aging, McGill University, Montreal, QC, Canada,* [3]*Alzheimer's Disease Research Unit, Douglas Mental Health University Institute, McGill University, Montreal, QC, Canada,* [4]*Department of Psychiatry, McGill University, Montreal, QC, Canada,* [5]*National Research Center for Dementia, Chosun University, Gwangju, South Korea,* [6]*Department of Biomedical Science, Chosun University, Gwangju, South Korea*

OPEN ACCESS

Edited by:
James H. Cole,
King's College London,
United Kingdom

Reviewed by:
Uicheul Yoon,
Catholic University of Daegu,
South Korea
Jungsu Oh,
Asan Medical Center, South Korea

***Correspondence:**
Kun Ho Lee
leekho@chosun.ac.kr
Boreom Lee
leebr@gist.ac.kr

†These authors have contributed equally to this work

Received: *09 November 2018*
Accepted: *10 January 2019*
Published: *11 February 2019*

Citation:
Qureshi MNI, Ryu S, Song J, Lee KH and Lee B (2019) Evaluation of Functional Decline in Alzheimer's Dementia Using 3D Deep Learning and Group ICA for rs-fMRI Measurements. Front. Aging Neurosci. 11:8. doi: 10.3389/fnagi.2019.00008

Purpose: To perform automatic assessment of dementia severity using a deep learning framework applied to resting-state functional magnetic resonance imaging (rs-fMRI) data.

Method: We divided 133 Alzheimer's disease (AD) patients with clinical dementia rating (CDR) scores from 0.5 to 3 into two groups based on dementia severity; the groups with very mild/mild (CDR: 0.5–1) and moderate to severe (CDR: 2–3) dementia consisted of 77 and 56 subjects, respectively. We used rs-fMRI to extract functional connectivity features, calculated using independent component analysis (ICA), and performed automated severity classification with three-dimensional convolutional neural networks (3D-CNNs) based on deep learning.

Results: The mean balanced classification accuracy was 0.923 ± 0.042 ($p < 0.001$) with a specificity of 0.946 ± 0.019 and sensitivity of 0.896 ± 0.077. The rs-fMRI data indicated that the medial frontal, sensorimotor, executive control, dorsal attention, and visual related networks mainly correlated with dementia severity.

Conclusions: Our CDR-based novel classification using rs-fMRI is an acceptable objective severity indicator. In the absence of trained neuropsychologists, dementia severity can be objectively and accurately classified using a 3D-deep learning framework with rs-fMRI independent components.

Keywords: dementia, progression assessment, imaging biomarkers, independent component analysis, neuroimaging, convolutional neural network

Abbreviations: 3D-CNN, three-dimensional convolutional neural networks; CDR, clinical dementia rating; CSF, cerebrospinal fluid; ICA, independent component analysis; rs-fMRI, resting-state functional magnetic resonance imaging; MCI, mild cognitive impairment.

INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia among dementia patients, with a 40%–60% prevalence (Ferri et al., 2005). It is a devastating illness and results in major cognitive and behavioral impairments. The established underlying mechanism is neurodegeneration, which is attributed to the accumulation of Aβ, hyperphosphorylation of tau proteins, and neuroinflammation (Leuner et al., 2007; Frautschy and Cole, 2010; Shadfar et al., 2015). Although molecular chemistry research on the mechanism of dementia has been conducted, numerous reports suggest structural and functional changes in the brain identified using neuroimaging (He et al., 2007; Solé-Padullés et al., 2009; Adlard et al., 2014). The assessment and treatment for patients with AD are multi-modal and are based on the stage of the illness. At each stage, the physician should alert and help the patients and their families to anticipate future symptoms and the related care that may be required. Although dementia symptoms can be controlled, slowing disease progression down is not a direct treatment for the pathophysiological mechanism of AD (Cummings and Fox, 2017). Given that most drugs currently used for treatment of AD patients act by enhancing cholinergic transmission and thus require viable synapses (DeKosky and Scheff, 1990; Terry and Buccafusco, 2003; Sarter and Parikh, 2005; Cacabelos, 2007), evaluation of the stage of dementia by experts is important for appropriate symptom control; additionally, the evaluation of viable synapse is important for determining the progression of the disease (DeKosky and Scheff, 1990; Scheff et al., 1990).

However, despite numerous neuroimaging studies, staging of dementia is generally based on past history and mental status examination by trained neuro-psychiatrists under the guidelines of the clinical dementia rating scale (CDR; Hughes et al., 1982). CDR helps the clinicians to rate the severity of AD and related disorders on a scale from 0 (normal) to 3 (severe stage) based on clinical interviews with a caregiver and the person with dementia. The areas that are coded are memory, orientation, judgment, problem-solving, community affairs, home, and hobbies. Despite the use of CDR, which is consensual among neuro-psychiatrists, and is based on extensive research and statistics to ensure the validity of the dementia severity rating, the diagnostic process mainly depends on the assessment of clinical symptoms. Furthermore, the diagnostic criteria of AD involves a substantial observation period and a reliable informant. In addition, it is too burdensome for a general doctor to use CDR (Perneczky et al., 2006). Also, CDR may have limitations in detecting early dementia (Rockwood et al., 2000; Schafer et al., 2004). Therefore, an additional tool for rating dementia severity is definitely required, and neuroimaging techniques may serve to complement the CDR scale.

Recently, resting-state functional connectivity is regarded as an important biomarker for AD. Several studies have reported that AD patients show decreased resting-state functional connectivity in the default mode network (DMN; Greicius et al., 2004; Hafkemeijer et al., 2012; Koch et al., 2012; Franciotti et al., 2013; Krajcovicova et al., 2014; Joo et al., 2016). Although atrophy was not observed, mild cognitive impairment (MCI) was associated with decreased functional connectivity of the medial temporal lobe or DMN region (Jin et al., 2012). Several resting-state functional magnetic resonance imaging (rs-fMRI) studies have addressed the issues of early detection, classification, and prediction in AD, MCI, normal patients, and subtypes of dementia. Previous reports have provided optimistic results for the classification of AD, MCI, and healthy normal aging individuals. Various approaches, such as independent component analysis (ICA; Fox et al., 2006; Dosenbach et al., 2007; Sylvester et al., 2009; Zhou et al., 2010), region of interest (Wang et al., 2006; Chen et al., 2011; Challis et al., 2015), graph theory (Supekar et al., 2008; Khazaee et al., 2015), multivoxel pattern analysis using machine learning (Mahmoudi et al., 2012), and multimodal (Dai et al., 2012; Dyrba et al., 2015) approaches have shown high performance (72%–94% accuracy). However, most prior studies have used datasets only from a single site/source, except for a study in which the AD neuroimaging initiative (ADNI) dataset was compared to their in-house dataset for validation of MCI/Normal classification algorithm (Suk et al., 2016). Therefore, the classification format of most previous studies strictly followed the form of the database. The ADNI dataset is aimed at early detection of AD, and related studies focus on classifying the normal patients, MCI, and early AD. Therefore, ADNI did not contain adequate numbers of severe-stage patients diagnosed with CDR 2 or 3 score (late AD).

ICA is an effective method for functional connectivity analysis of brain imaging data (Lu and Rajapakse, 2006; Rajapakse and Zhou, 2007; Brier et al., 2012). Previously, numerous studies have reported greater functional connectivity in the salience (SAL) of patients with mild dementia (primarily CDR 1) than in normal individuals (Fox et al., 2006; Dosenbach et al., 2007; Sylvester et al., 2009; Zhou et al., 2010). In contrast, functional connectivity increments of the SAL were seen at levels between CDR 0 and CDR 0.5, which implicates a reduced correlation at CDR 1. This difference depends on the method used to acquire the independent components (Brier et al., 2012). In the past, the ICA components were reviewed by trained clinicians for the selection of meaningful components (Oh et al., 2017). Currently, ICA components can be automatically selected using highly advanced algorithms (Beckmann et al., 2009; Filippini et al., 2009). On applying these algorithms, we can consistently and automatically select the ICA components in classification studies.

Deep learning has gained enormous attention (Gal and Ghahramani, 2016; Amiri et al., 2018) in the last few years. The recent advances in machine learning in terms of image understanding have led to great advances with respect to identifying, classifying, and quantifying patterns of medical images, especially using deep learning. In particular, the utilization of hierarchical functional representations learned solely with data, instead of manually created features that are designed based on domain-specific knowledge is at the core of the progress (Raju et al., 2017; Shen et al., 2017; Amiri et al., 2018). Previous studies have reported that the classification of dementia, MCI, and normal individuals can be performed

automatically using deep learning and multimodal data including neuroimaging data or biological measures from cerebrospinal fluid (CSF; Suk and Shen, 2013; Liu et al., 2015; Suk et al., 2015). Automated diagnostics using multimodal neuroimaging data have the advantage of utilizing all information, and demonstrate the potential to improve diagnostic accuracy. However, the process is highly complex and requires additional computational resources. Therefore, it would be preferable to obtain acceptable accuracy with only unimodal data.

Three-dimensional convolutional neural network (3D-CNN) in deep learning is a supervised learning framework and is enabled to distinguish training data similar to the visual processing of the human eye (Ji et al., 2013). While these networks have been used specifically for visual recognition in the 2D domain over the last few years by researchers in visual computing and artificial intelligence research, it is unlikely that 3D-CNN was used for volumetric neuroimaging data classification and prediction. The novelty of this study is that 3D ICA data were used as input for the 3D-CNN model. Considering that previous studies have shown that group ICA features have the potential to discriminate dementia severity, we classified the severity of dementia using 3D deep learning with group ICA input.

Despite its clinical importance, the severity estimation of AD using image data was not conducted by any researcher at all, except for one report that characterizes five resting state networks of CDR 0.5 and 1 (Brier et al., 2012). Therefore, a major novel feature of our research is the automatic classification of AD into two groups of disease severity (very mild and mild vs. moderate and severe).

To propose an alternative method to complement the CDR scale in the evaluation of AD, we hypothesized that the functional connectivity changes according to the stage of AD will be observed in the rs-fMRI, and the severity of AD could be classified using 3D-CNN.

MATERIALS AND METHODS

Dataset

This dataset was a part of a large cohort enrolled at National Dementia Research Center, Chosun University, Gwangju, South Korea. Each subject provided written informed consent before the data collection. The data acqusistion was approved by the institutional review board of the Chosun University Hospital, Gwangju, South Korea (IRB number 2013-12-018).

The demographics of the participants are shown in **Table 1**. CDR is a categorical variable. To better estimate the decline of resting-state functional connectivity with increasing AD severity, we allocated the labeled data into two groups. Group 1 includes very mild to mild (CDR 0.5 and 1.0) and group 2 includes moderate to severe (CDR 2.0–3.0) patients.

TABLE 1 | Subject demographics.

	Very mild to mild AD (*n* = 77; 30 F/47 M)	Moderate to severe AD (*n* = 49; 32 F/17 M)	*p*-value
Age (years)	73.57 ± 6.49	73.61 ± 4.76	0.160
Education (score)	10.09 ± 4.95	6.79 ± 4.54	0.227
MMSE (score)	23.84 ± 3.90	15.49 ± 4.87	0.09
CDR (score)	0.71 ± 0.25	2.08 ± 0.28	0.001*

The p-value was computed by applying the t-test to the clinical dementia rating (CDR) scores.

Resting-State fMRI Data Acquisition

All the participants were scanned with a Siemens Skyra 3.0-Tesla scanner. A 2D EPI MR acquisition type was used with the following parameters: TR/TE = 3,000/30 ms, flip angle = 90°, FOV = 240 × 240 mm, voxel size = 3.75 × 3.75 × 3.75, spacing between slices = 4.8 mm, number of echoes = 1, imaging frequency = 123.206 Hz, slice acquisition order = ascending (bottom-up), direction = 'Transverse > Coronal (2.6) > Saggital (1.7)', pixel bandwidth = 3440, inplane phase encoding direction = 'ROW', number of phase encoding steps = 63, echo train length = 31° sampling = 100° phase field of view = 100, variable flip angle flag = 'N', and SAR = 0.0778.

Preprocessing of the Resting-State fMRI Data

The rs-fMRI data was pre-processed with FMRIB Software Library (FSL[1]) version 6.0. Standard preprocessing routines were applied with motion correction, slice timing correction, spatial smoothing with 6 mm full width half maximum Gaussian kernel, temporal filtering, and thereafter each subject's functional data were co-registered to its corresponding structural image. Subsequently, for acquiring the group ICA based connectivity measures, FSL Multivariate Exploratory Linear Optimized Decomposition into Independent Components (MELODIC) version 3.14 was utilized to perform a single-session ICA. The number of independent components was set as 30 (Qureshi et al., 2017). We used variance normalization and thresholded the independent component maps with an alternative hypothesis test that was based on the fitting of a Gaussian/gamma mixture model to the distributions of the voxel intensities within the spatial maps and controlling the local false-discovery rate at $p < 0.5$. The set of spatial maps from the group-average analysis was used to generate subject-specific versions of the spatial maps, and associated time-series, using dual regression (Beckmann et al., 2005, 2009). First, for each subject, the group-average set of spatial maps is regressed (as spatial regressors in a multiple regression) into the subject's 4D space-time dataset (Oh et al., 2017; Qureshi et al., 2017). This results in a set of subject-specific time-series, one per group-level spatial map. Next, those time series are regressed (as temporal regressors, again in a multiple regression) into the same 4D dataset, resulting in a set of subject-specific 3D spatial maps, one per group-level. We then tested for group differences, using FSL's randomized permutation-testing tool (Smith et al., 2004). Among the 30 independent components, 15 were classified as noise and/or artifacts using the automated clustering tool of FSLNets[2]. Besides the automated selection, these components were also validated by visual inspection by an experienced clinical neurologist, similar to the procedure used in our

[1] www.fmrib.ox.ac.uk/fsl

[2] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSLNets

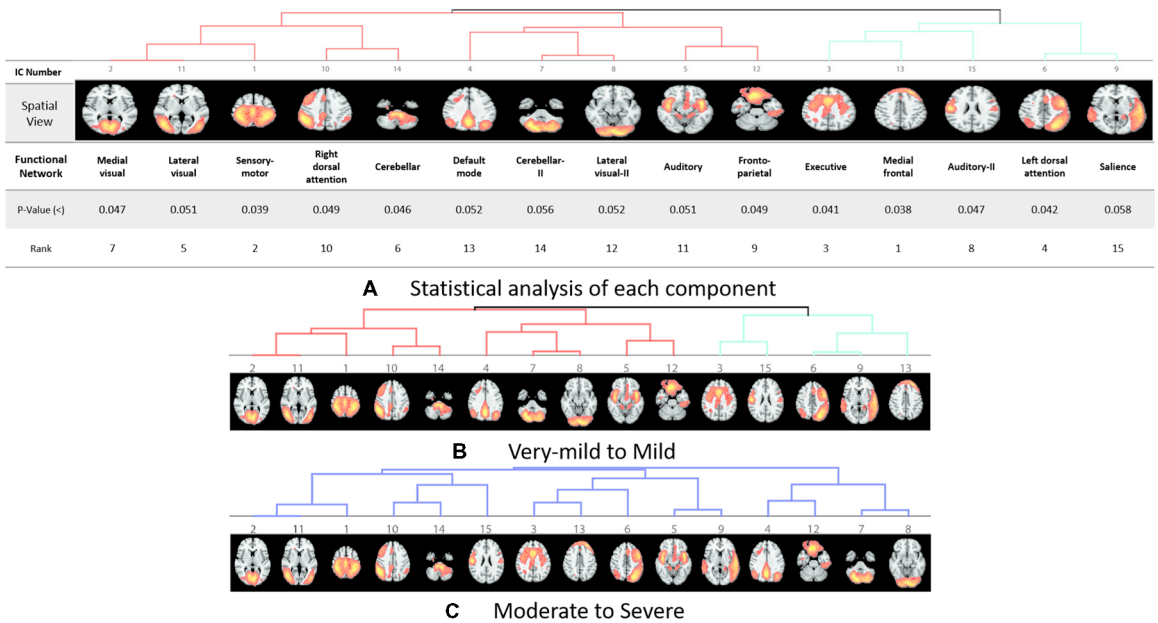

FIGURE 1 | Dendrograms of the selected useful independent component-based functional networks using automated clustering. **(A)** Statistical analysis of each component. **(B)** Dendrogram of very-mild and mild groups, two major divisions are shown in red and green color. **(C)** Dendrogram of moderate and severe groups, no significant division was found among the functional networks according to the clinical dementia rating (CDR) level.

previous studies (Qureshi et al., 2017). **Figure 1A** depicts the selected 15 components. It represents the well-known resting-state functional networks including the DMN, sensorimotor network, medial and lateral visual network, left and right dorsal attention network, central executive network, cerebellar network, salience network, limbic network, auditory network, and frontal networks.

Features

We used the 3D volumetric images of these selected functional networks for the classification between the CDR low and CDR high groups. These 3D images were acquired by performing dual regression (Beckmann et al., 2009) on the group ICA result.

Deep Learning and 3D-CNN Framework

We used a 3D-CNN based deep learning classification framework in this study. This framework was implemented on the TensorFlow library version 1.5 with Nvidia Geforce GTX 1080Ti graphical processing unit (GPU) support. For the training model, we used the Adam optimizer with a learning rate of 0.001, epsilon value was set at 0.1, and minimal cost was used. Since the size of the dataset was relatively small for deep learning, to avoid model overfit, we used ten-fold cross-validation in this study to report the mean accuracy of the model. A modified version of VGG-Net classification framework was used in this study. Specifically, we added batch normalization layers in the convolution layer. A dropout rate of 0.7 was used in the fully connected layers. The batch size was set at 12 and 50 epochs were used. The parameters including learning rate, epsilon value, dropout rate, batch size, and epoch size were optimized using the following ranges. For epsilon, we tunned it in the range of [0.1 : 0.05 : 1], for learning rate, we tunned it in the logarithmic range of [1, 0.1, 0.01, 0.001, 0.0001, and 0.00001], for the dropout rate, we tunned it in the range [0.1 : 0.05 : 1], for the batch size, we optimized it by the maximum available GPU memory, and the number of epochs were tunned in the range of [10 : 1 : 200]. To the best of our knowledge, CNN is the only deep learning framework that learn from 3D input, therefore no other deep learning architectures were tested during this study. **Figure 2** depicts the complete architecture of our 3D-CNN deep classification framework. Details of the model are given in **Table 2**.

Significance Testing

For assessing the statistical significance of the results, we performed the permutation test on the classification accuracies and permuted the labels of test data of each of the 10 folds 1,000 times to get the probability of successful classification with a higher score than the actual test labels.

RESULTS

Our results suggest that CDR level can be used as a good discriminatory predictor of the dementia stages. We achieved a mean balanced test accuracy of 92.30% in a ten-fold cross validation experiment using the 3D-CNN algorithm.

Classification

We achieved an optimistic 10-fold cross-validated classification accuracy. Since the dataset was not balanced, we also computed the balanced accuracy to remove any bias present in the result due to unbalanced data. **Table 3** shows all the

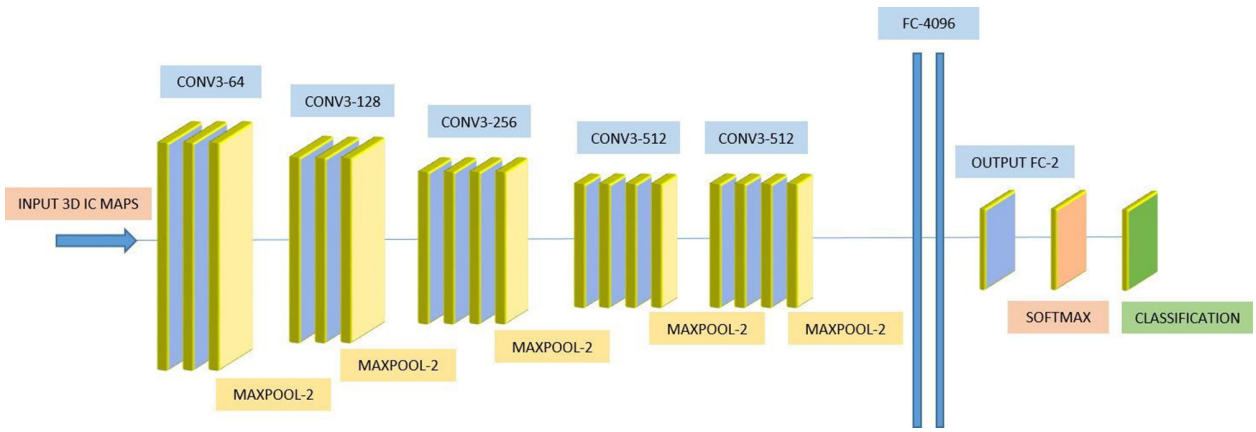

FIGURE 2 | VGG-Net based three-dimensional convolutional neural network (3D-CNN) architecture.

performance evaluation measures in the data including the test accuracy, train accuracy, specificity, sensitivity, and balanced accuracy.

Statistical Significance

Statistically, this result has very high significance with $p < 0.001$ for all the 10-folds of the classification experiment. The significance measure through permutation testing were computhed as the *p*-values as mentioned in **Table 3** for each fold of the cross-validation.

Clinical Significance

These results suggest that CDR-based novel classification of rs-fMRI can be accepted as an objective severity index. **Table 4** shows the ranking of each functional network as the features of a deep learning framework based on the unpaired *t*-test. The uncorrected *p*-value revealed the component's significance. **Figure 3** shows the connectogram of the selected networks.

DISCUSSION

To the best of our knowledge, this is a pioneering study to classify the severity of dementia using rs-fMRI and 3D-CNN deep learning architecture rather than a 1D time-series information. Because the assessment of symptoms of patients with AD is important for appropriate treatment, the automatic classification of AD of the two groups of disease severity has important contributions for clinical practice.

TABLE 2 | Details of the three-dimensional convolutional neural network (3D-CNN) architecture.

Layer	Feature Map	Stride	Kernel	Activation structure
Convolution	64	1 × 1 × 1	3 × 3 × 3	Conv
Convolution	64	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Maxpool		2 × 2 × 2	2 × 2 × 2	
Convolution	128	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Convolution	128	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Maxpool		1 × 1 × 1	2 × 2 × 2	
Convolution	256	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Convolution	256	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Convolution	256	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Maxpool		2 × 2 × 2	2 × 2 × 2	
Convolution	512	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Convolution	512	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Convolution	512	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Maxpool		2 × 2 × 2	2 × 2 × 2	
Convolution	512	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Convolution	512	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Convolution	512	1 × 1 × 1	3 × 3 × 3	Batchnorm+ReLU+Conv
Maxpool		1 × 1 × 1	2 × 2 × 2	
Fully Connected	4096	Dropout rate 0.7		ReLU
Fully Connected	4096	Dropout rate 0.7		ReLU
Output				
Fully Connected	2			
Softmax				
Classification Layer		Argmax		

TABLE 3 | Classification accuracy using 10-fold cross-validation.

Fold	Train Acc (%)	Test Acc (%)	*p*-value	AUC	Specificity (%)	Sensitivity (%)	BAC
1	99.44	95.83	<0.001	0.9936	0.9421	0.9871	0.9647
2	99.72	95.31	<0.001	0.9832	0.9561	0.9294	0.9428
3	99.94	91.66	<0.001	0.9838	0.9162	0.8864	0.9161
4	99.94	88.02	<0.001	0.9649	0.9320	0.8021	0.8671
5	99.94	91.14	<0.001	0.9767	0.9532	0.8587	0.9059
6	99.94	95.83	<0.001	0.9934	0.9734	0.9419	0.9577
7	99.04	94.79	<0.001	0.9809	0.9483	0.9398	0.9441
8	99.88	83.85	<0.001	0.9765	0.9456	0.7383	0.8419
9	99.61	92.18	<0.001	0.9740	0.9231	0.9146	0.9189
10	99.72	96.88	<0.001	0.9936	0.9739	0.9642	0.9690
Mean ± SD	99.72 ± 0.29	92.55 ± 4.11		0.982 ± 0.009	0.946 ± 0.019	0.896 ± 0.077	0.923 ± 0.042

TABLE 4 | Statistical analysis of each component.

Component name	Component number	uncorrected *p*-value (<)	Rank
Sensory-motor network	1	0.038933	2
Medial visual-related network	2	0.046845	7
Executive control network	3	0.040517	3
Default mode network	4	0.052597	13
Auditory related network	5	0.051502	11
Left dorsal attention network	6	0.042201	4
Cerebellar network	7	0.056729	14
Lateral visual-related network	8	0.052143	12
Salience network	9	0.058392	15
Right dorsal attention network	10	0.049393	10
Lateral visual-related network-II	11	0.043619	5
Fronto-parietal network	12	0.049307	9
Medial frontal network	13	0.038227	1
Cerebellar network-II	14	0.045902	6
Auditory related network-II	15	0.047154	8

There are previous studies on automated diagnosis using deep learning and multimodal neuroimaging data involving the CSF and laboratory assessments. Among these, there are numerous studies that classified dementia, MCI, and healthy individuals (Suk and Shen, 2013; Liu et al., 2015; Suk et al., 2015). It may be helpful to analyze structural MRI changes in distinguishing between normal patients, MCI, and AD. However, since structural changes are more likely to have progressed beyond a certain level, structural MRI may act as a confounding factor when considering individual differences. CSF studies may be helpful in assessing severity. To acquire CSF samples, we perform an invasive procedure, which is a lumbar puncture. However, considering the enviornment of out patient departments in Korean hospitals, it is diffciult to perform invasive procedures. Overall, if cost-effectiveness was taken into account, it would be best that the severity was determined using only noninvasive rs-fMRI. If only rs-fMRI was used, the imaging time could be less than a few minutes and may prove effective in clinical management.

Only one study reported the characteristics of five resting state networks of CDR score from 0.5 to 1 (Brier et al., 2012).

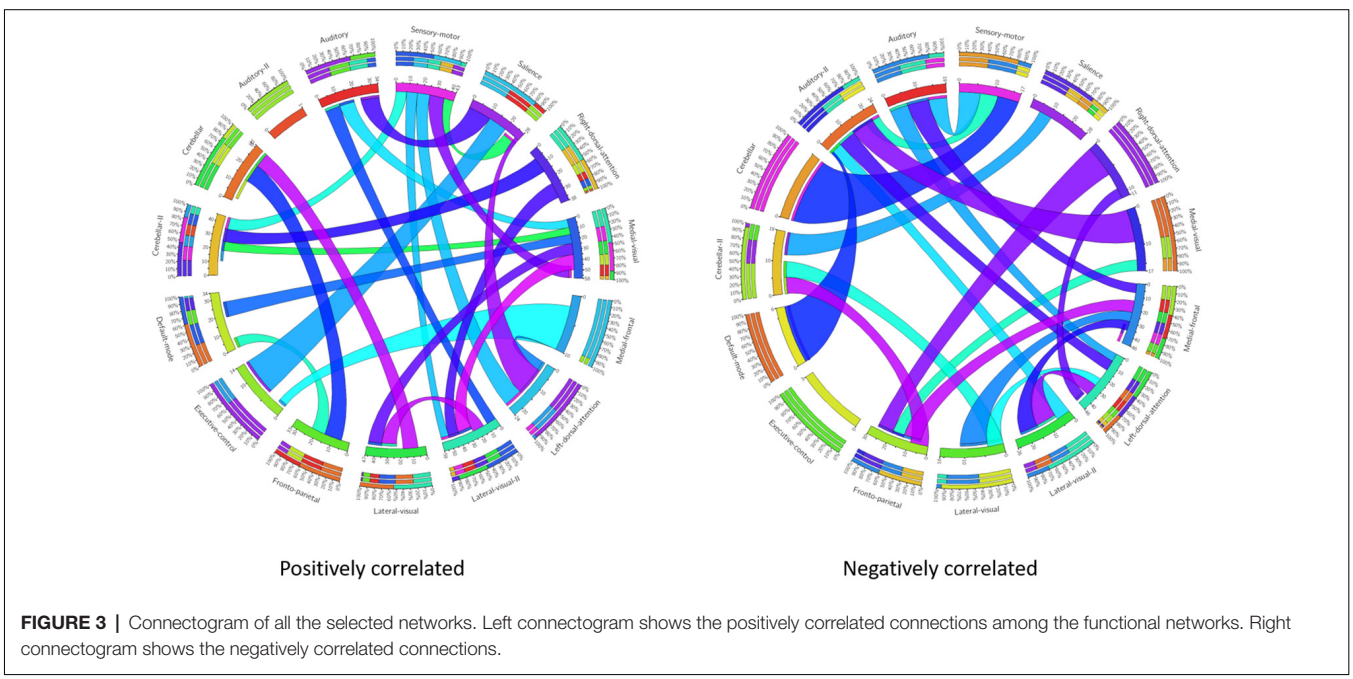

FIGURE 3 | Connectogram of all the selected networks. Left connectogram shows the positively correlated connections among the functional networks. Right connectogram shows the negatively correlated connections.

This report provided clues to the discriminatory potential of group ICA features that could contribute to the classification of dementia severity. However, no study has been conducted on patients with CDR scores of 2 or 3 with ICA as features, which were classified automatically from noise using FSLNet and deep learning structure. Therefore, the major contribution of our research is the automatic classification of AD into two groups of disease severity (very mild and mild vs. moderate and severe).

Our results showed a mean test accuracy of 92.30% in a 10-fold cross validation experiment using the 3D-CNN algorithm. We believe that a deep neural network constitutes the optimal classification weight through iterative learning, but the extent of contribution of the ICA component of deep learning architecture to the algorithm is not known. To reveal the black box of 3D-CNN, we also compared each component between very mild/mild vs. moderate/ severe patients. Previous studies have reported that the DMN is the most significant different functional network between normal patients and MCI and dementia (Wang et al., 2006; Jin et al., 2012; Koch et al., 2012), and the salience network had differences between CDR 0.5 and 1 (Fox et al., 2006; Dosenbach et al., 2007; Sylvester et al., 2009; Zhou et al., 2010). Interestingly, our result showed that the medial frontal, sensory-motor, executive control, left dorsal attention, lateral visual-related, cerebellar, medial visual-related, auditory-related, frontoparietal, and right dorsal attention networks have high ranks and statistical differences. After the onset of dementia, functional connectivity seems to be observed in an altered way. We assumed that those networks have more influence on our classifier. Although DMN and salience network do not have enough statistical significance, the combination of the information from various components and their relationship including functional connectivity may contribute to the classification algorithm. **Figures 1**, **2** show the relationships among the components. Red color represents positive correlations and blue color represents negative correlation among the components. These associations represent the activity of each component, and there were no significant differences between the two groups, which is also shown in **Table 4**. Even in case of subtle differences, with deep learning these can be utilized to extract features to render the weights more suitable.

Research on drug development for AD has not been able to improve drug-based treatments, in spite of the recently advanced understanding of the molecular-cellular biology of the disease (De Strooper, 2014; Gauthier et al., 2016). Although, there may be numerous reasons for the failure of new drug development, as the stage of dementia differs from patient to patient, it is difficult to evaluate the response to symptoms alone. In addition, dementia could be a confounding factor due to the differences in the characteristics of individuals including genomic, proteomic, and metabolomic cascades. A previous study reported that current trials have focused on clinical efficacy and not on the rigorous testing of the putative mechanisms of disease (Becker et al., 2014). Considering that the central cholinergic deficit in AD is the consequence of neurodegeneration, the imaging method of measuring viable synapses is appropriate for evaluating drug responses. Because fMRI measures the function of the brain through the blood oxygen level dependent technique, it may help to compensate for the weaknesses of drug efficacy assessment through symptoms. Our classification algorithm based unimodal rs-fMRI extracts features from the degeneration of the functional connectivity in dementia. During the evaluation of drug response or behavioral therapy according to the stage and symptoms of AD, it would be helpful to investigate the recovery of functional connectivity objectively.

The novelty of our study is that we analyzed the severity of dementia, although our study also has limitations. We used our dataset to create a 3D-CNN classifier, but we could not perform the verification procedure with other datasets. Because of the ADNI dataset, which has been widely used in previous dementia studies, we could focus on early stage dementia detection; and the numbers of late-stage dementia patients were not adequate for comparison. It is necessary to apply our algorithm to other datasets with adequate numbers of patients with late-stage dementia.

Another limitation is due to the characteristics of deep learning. A total of 15 ICAs were selected as input for deep learning, but it is difficult to determine the precise effect on the neural network. To overcome this limitation, we statistically analyzed the differences of ICA between the two groups.

Another limition of the present study is in terms of the limited number of subjects, however, it is inappropriate to apply standard data augmentation approaches on the neuroimaging data to increase the number of training samples. We believe that the introduction of any type of synthesized data in training phase can significantly bias the learning process. In addition, the signal to noise ratio in fMRI data is relatively small therefore it is very difficult to apply deep learning to the raw data. A major advantage of using ICA is the removal of artifacts because they very much look like the BOLD signal in raw data.

One of the most important aspects of this research is the use of neuroimaging to predict the progression of diseases that humans can not predict, especially for the subjects with MCI who progress to dementia as compared to those who do not progress to dementia in the future. However, we cannot represent it in our present study. The classification task in this research has a limitation because the data labels used in this study are based on contemporary clinical evaluations. In addition, classifying current disease status is important but clinically, predicting the progression from MCI to dementia and classifying severity in dementia is more important for proper and appropriate treatment, and also prediction from MCI to dementia and current severity classification can have a decisive impact on prognosis. Taking all of these measures into account, our analysis can be considered as a clinically relevant study involving future outcomes. In the future, we will also perform an advanced study to predict the progression from MCI to dementia using

biomarker-based serial labeled data and domain transfer learning methods.

In conclusion, our study suggests that our novel classifier using rs-fMRI is acceptable as an objective severity indicator complementing the CDR scale in the evaluation of AD. In the absence of trained neurologists, we can classify the dementia severity objectively and accurately using 3D-deep learning. Our application and classification algorithm would be an aid for observing the regeneration of functional connectivity due to drug treatment according to the stage and symptoms of AD in the future.

DATA AVAILABILITY

The datasets for this study will not be made publicly available because this study cohort is not open for public use.

AUTHOR CONTRIBUTIONS

MQ and SR have equally contributed in this work. MQ developed the whole idea of this work. SR made the clinical representation of the results. JS helped in the design of deep learning framework. KL and BL supervised the research.

FUNDING

This work was supported by GIST Research Institute (GRI) grant funded by the GIST in 2019; the Bio and Medical Technology Development Program of the NRF funded by the Korean government, MSIT (Grant No. 2016M3A9E9941946); and the Original Technology Research Program for Brain Science of the NRF funded by the Korean government, MSIT (NRF-2014M3C7A1046041).

REFERENCES

Adlard, P. A., Tran, B. A., Finkelstein, D. I., Desmond, P. M., Johnston, L. A., Bush, A. I., et al. (2014). A review of β-amyloid neuroimaging in Alzheimer's disease. *Front. Neurosci.* 8:327. doi: 10.3389/fnins.2014.00327

Amiri, S., Mahjoub, M. A., and Rekik, I. (2018). "Bayesian network and structured random forest cooperative deep learning for automatic multi-label brain tumor segmentation," in *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018)—Volume 2,* 183–190.

Becker, R. E., Greig, N. H., Giacobini, E., Schneider, L. S., and Ferrucci, L. (2014). A new roadmap for drug development for Alzheimer's disease. *Nat. Rev. Drug Discov.* 13:156. doi: 10.1038/nrd3842-c2

Beckmann, C. F., DeLuca, M., Devlin, J. T., and Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 360, 1001–1013. doi: 10.1098/rstb.2005.1634

Beckmann, C. F., Mackay, C. E., Filippini, N., and Smith, S. M. (2009). Group comparison of resting-state FMRI data using multi-subject ICA and dual regression. *Neuroimage* 47:S148. doi: 10.1016/s1053-8119(09)71511-3

Brier, M. R., Thomas, J. B., Snyder, A. Z., Benzinger, T. L., Zhang, D., Raichle, M. E., et al. (2012). Loss of intranetwork and internetwork resting state functional connections with Alzheimer's disease progression. *J. Neurosci.* 32, 8890–8899. doi: 10.1523/JNEUROSCI.5698-11.2012

Cacabelos, R. (2007). Donepezil in Alzheimer's disease: from conventional trials to pharmacogenetics. *Neuropsychiatr. Dis. Treat.* 3:303.

Challis, E., Hurley, P., Serra, L., Bozzali, M., Oliver, S., and Cercignani, M. (2015). Gaussian process classification of Alzheimer's disease and mild cognitive impairment from resting-state fMRI. *Neuroimage* 112, 232–243. doi: 10.1016/j.neuroimage.2015.02.037

Chen, G., Ward, B. D., Xie, C., Li, W., Wu, Z., Jones, J. L., et al. (2011). Classification of Alzheimer disease, mild cognitive impairment and normal cognitive status with large-scale network analysis based on resting-state functional MR imaging. *Radiology* 259, 213–221. doi: 10.1148/radiol.10100734

Cummings, J., and Fox, N. (2017). Defining disease modifying therapy for Alzheimer's disease. *J. Prev. Alzheimers Dis.* 4, 109–115. doi: 10.14283/jpad.2017.12

Dai, Z., Yan, C., Wang, Z., Wang, J., Xia, M., Li, K., et al. (2012). Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3). *Neuroimage* 59, 2187–2195. doi: 10.1016/j.neuroimage.2011.10.003

De Strooper, B. (2014). Lessons from a failed γ-secretase Alzheimer trial. *Cell* 159, 721–726. doi: 10.1016/j.cell.2014.10.016

DeKosky, S. T., and Scheff, S. W. (1990). Synapse loss in frontal cortex biopsies in Alzheimer's disease: correlation with cognitive severity. *Ann. Neurol.* 27, 457–464. doi: 10.1002/ana.410270502

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A. T., et al. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proc. Natl. Acad. Sci. U S A* 104, 11073–11078. doi: 10.1073/pnas.0704320104

Dyrba, M., Grothe, M., Kirste, T., and Teipel, S. J. (2015). Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum. Brain Mapp.* 36, 2118–2131. doi: 10.1002/hbm.22759

Ferri, C. P., Prince, M., Brayne, C., Brodaty, H., Fratiglioni, L., Ganguli, M., et al. (2005). Global prevalence of dementia: a Delphi consensus study. *Lancet* 366, 2112–2117. doi: 10.1016/S0140-6736(05)67889-0

Filippini, N., MacIntosh, B. J., Hough, M. G., Goodwin, G. M., Frisoni, G. B., Smith, S. M., et al. (2009). Distinct patterns of brain activity in young carriers of the APOE-ϵ4 allele. *Proc. Natl. Acad. Sci. U S A* 106, 7209–7214. doi: 10.1073/pnas.0811879106

Fox, M. D., Corbetta, M., Snyder, A. Z., Vincent, J. L., and Raichle, M. E. (2006). Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems. *Proc. Natl. Acad. Sci. U S A* 103, 10046–10051. doi: 10.1073/pnas.0606682103

Franciotti, R., Falasca, N. W., Bonanni, L., Anzellotti, F., Maruotti, V., Comani, S., et al. (2013). Default network is not hypoactive in dementia with fluctuating cognition: an Alzheimer disease/dementia with Lewy bodies comparison. *Neurobiol. Aging* 34, 1148–1158. doi: 10.1016/j.neurobiolaging.2012.09.015

Frautschy, S. A., and Cole, G. M. (2010). Why pleiotropic interventions are needed for Alzheimer's disease. *Mol. Neurobiol.* 41, 392–409. doi: 10.1007/s12035-010-8137-1

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning* (New York, NY: ACM), 1050–1059.

Gauthier, S., Albert, M., Fox, N., Goedert, M., Kivipelto, M., Mestre-Ferrandiz, J., et al. (2016). Why has therapy development for dementia failed in the last two decades? *Alzheimers Dement.* 12, 60–64. doi: 10.1016/j.jalz.2015.12.003

Greicius, M. D., Srivastava, G., Reiss, A. L., and Menon, V. (2004). Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI. *Proc. Natl. Acad. Sci. U S A* 101, 4637–4642. doi: 10.1073/pnas.0308627101

Hafkemeijer, A., van der Grond, J., and Rombouts, S. A. R. B. (2012). Imaging the default mode network in aging and dementia. *Biochim. Biophys. Acta* 1822, 431–441. doi: 10.1016/j.bbadis.2011.07.008

He, Y., Wang, L., Zang, Y., Tian, L., Zhang, X., Li, K., et al. (2007). Regional coherence changes in the early stages of Alzheimer's disease: a combined structural and resting-state functional MRI study. *Neuroimage* 35, 488–500. doi: 10.1016/j.neuroimage.2006.11.042

Hughes, C. P., Berg, L., Danziger, W. L., Coben, L. A., and Martin, R. L. (1982). A new clinical scale for the staging of dementia. *Br. J. Psychiatry* 140, 566–572. doi: 10.1192/bjp.140.6.566

Ji, S., Yang, M., and Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 221–2231. doi: 10.1109/tpami.2012.59

Jin, M., Pelak, V. S., and Cordes, D. (2012). Aberrant default mode network in subjects with amnestic mild cognitive impairment using resting-state functional MRI. *Magn. Reson. Imaging* 30, 48–61. doi: 10.1016/j.mri.2011.07.007

Joo, S. H., Lim, H. K., and Lee, C. U. (2016). Three large-scale functional brain networks from resting-state functional MRI in subjects with different levels of cognitive impairment. *Psychiatry Investig.* 13, 1–7. doi: 10.4306/pi.2016.13.1.1

Khazaee, A., Ebrahimzadeh, A., and Babajani-Feremi, A. (2015). Identifying patients with Alzheimer's disease using resting-state fMRI and graph theory. *Clin. Neurophysiol.* 126, 2132–2141. doi: 10.1016/j.clinph.2015.02.060

Koch, W., Teipel, S., Mueller, S., Benninghoff, J., Wagner, M., Bokde, A. L. W., et al. (2012). Diagnostic power of default mode network resting state fMRI in the detection of Alzheimer's disease. *Neurobiol. Aging* 33, 466–478. doi: 10.1016/j.neurobiolaging.2010.04.013

Krajcovicova, L., Marecek, R., Mikl, M., and Rektorova, I. (2014). Disruption of resting functional connectivity in Alzheimer's patients and at-risk subjects. *Curr. Neurol. Neurosci. Rep.* 14:491. doi: 10.1007/s11910-014-0491-3

Leuner, K., Hauptmann, S., Abdel-Kader, R., Scherping, I., Keil, U., Strosznajder, J. B., et al. (2007). Mitochondrial dysfunction: the first domino in brain aging and Alzheimer's disease? *Antioxid. Redox Signal.* 9, 1659–1675. doi: 10.1089/ars.2007.1763

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., et al. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. *IEEE Trans. Biomed. Eng.* 62, 1132–1140. doi: 10.1109/tbme.2014.2372011

Lu, W., and Rajapakse, J. C. (2006). ICA with reference. *Neurocomputing* 69, 2244–2257. doi: 10.1016/j.neucom.2005.06.021

Mahmoudi, A., Takerkart, S., Regragui, F., Boussaoud, D., and Brovelli, A. (2012). Multivoxel pattern analysis for FMRI data: a review. *Comput. Math. Methods Med.* 2012:961257. doi: 10.1155/2012/961257

Oh, J., Chun, J.-W., Kim, E., Park, H.-J., Lee, B., and Kim, J.-J. (2017). Aberrant neural networks for the recognition memory of socially relevant information in patients with schizophrenia. *Brain Behav.* 7:e00602. doi: 10.1002/brb3.602

Perneczky, R., Wagenpfeil, S., Komossa, K., Grimmer, T., Diehl, J., and Kurz, A. (2006). Mapping scores onto stages: mini-mental state examination and clinical dementia rating. *Am. J. Geriatr. Psychiatry* 14, 139–144. doi: 10.1097/01.jgp.0000192478.82189.a8

Qureshi, M. N. I., Oh, J., Cho, D., Jo, H. J., and Lee, B. (2017). Multimodal discrimination of schizophrenia using hybrid weighted feature concatenation of brain functional connectivity and anatomical features with an extreme learning machine. *Front. Neuroinform.* 11:59. doi: 10.3389/fninf.2017.00059

Rajapakse, J. C., and Zhou, J. (2007). Learning effective brain connectivity with dynamic Bayesian networks. *Neuroimage* 37, 749–760. doi: 10.1016/j.neuroimage.2007.06.003

Raju, M., Pagidimarri, V., Barreto, R., Kadam, A., Kasivajjala, V., and Aswath, A. (2017). Development of a deep learning algorithm for automatic diagnosis of diabetic retinopathy. *Stud. Health Technol. Inform.* 245, 559–563.

Rockwood, K., Strang, D., MacKnight, C., Downer, R., and Morris, J. C. (2000). Interrater reliability of the Clinical Dementia Rating in a multicenter trial. *J. Am. Geriatr. Soc.* 48, 558–559. doi: 10.1111/j.1532-5415.2000.tb05004.x

Sarter, M., and Parikh, V. (2005). Choline transporters, cholinergic transmission and cognition. *Nat. Rev. Neurosci.* 6, 48–56. doi: 10.1038/nrn1588

Schafer, K. A., Tractenberg, R. E., Sano, M., Mackell, J. A., Thomas, R. G., Gamst, A., et al. (2004). Reliability of monitoring the clinical dementia rating in multicenter clinical trials. *Alzheimer Dis. Assoc. Disord.* 18, 219–222.

Scheff, S. W., DeKosky, S. T., and Price, D. A. (1990). Quantitative assessment of cortical synaptic density in Alzheimer's disease. *Neurobiol. Aging* 11, 29–37. doi: 10.1016/0197-4580(90)90059-9

Shadfar, S., Hwang, C. J., Lim, M.-S., Choi, D.-Y., and Hong, J. T. (2015). Involvement of inflammation in Alzheimer's disease pathogenesis and therapeutic potential of anti-inflammatory agents. *Arch. Pharm. Res.* 38, 2106–2119. doi: 10.1007/s12272-015-0648-x

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* 23, S208–S219. doi: 10.1016/j.neuroimage.2004.07.051

Solé-Padullés, C., Bartrés-Faz, D., Junqué, C., Vendrell, P., Rami, L., Clemente, I. C., et al. (2009). Brain structure and function related to cognitive reserve variables in normal aging, mild cognitive impairment and Alzheimer's disease. *Neurobiol. Aging* 30, 1114–1124. doi: 10.1016/j.neurobiolaging.2007.10.008

Suk, H.-I., Lee, S.-W., and Shen, D. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct. Funct.* 220, 841–859. doi: 10.1007/s00429-013-0687-3

Suk, H.-I., and Shen, D. (2013). Deep learning-based feature representation for AD/MCI classification. *Med. Image Comput. Comput. Assist. Interv.* 16, 583–590. doi: 10.1007/978-3-642-40763-5_72

Suk, H.-I., Wee, C.-Y., Lee, S.-W., and Shen, D. (2016). State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *Neuroimage* 129, 292–307. doi: 10.1016/j.neuroimage.2016.01.005

Supekar, K., Menon, V., Rubin, D., Musen, M., and Greicius, M. D. (2008). Network analysis of intrinsic functional brain connectivity in Alzheimer's disease. *PLoS Comput. Biol.* 4:e1000100. doi: 10.1371/journal.pcbi.1000100

Sylvester, C. M., Shulman, G. L., Jack, A. I., and Corbetta, M. (2009). Anticipatory and stimulus-evoked blood oxygenation level-dependent modulations related to spatial attention reflect a common additive signal. *J. Neurosci.* 29, 10671–10682. doi: 10.1523/JNEUROSCI.1141-09.2009

Terry, A. V. Jr., and Buccafusco, J. J. (2003). The cholinergic hypothesis of age and Alzheimer's disease-related cognitive deficits: recent challenges and their implications for novel drug development. *J. Pharmacol. Exp. Ther.* 306, 821–827. doi: 10.1124/jpet.102.041616

Wang, K., Jiang, T., Liang, M., Wang, L., Tian, L., Zhang, X., et al. (2006). Discriminative analysis of early Alzheimer's disease based on two intrinsically anti-correlated networks with resting-state fMRI. *Med. Image Comput. Comput. Assist. Interv.* 9, 340–347. doi: 10.1007/11866763_42

Wang, L., Zang, Y., He, Y., Liang, M., Zhang, X., Tian, L., et al. (2006). Changes in hippocampal connectivity in the early stages of Alzheimer's disease: evidence from resting state fMRI. *Neuroimage* 31, 496–504. doi: 10.1016/j.neuroimage.2005.12.033

Zhou, J., Greicius, M. D., Gennatas, E. D., Growdon, M. E., Jang, J. Y., Rabinovici, G. D., et al. (2010). Divergent network connectivity changes in behavioural variant frontotemporal dementia and Alzheimer's disease. *Brain* 133, 1352–1367. doi: 10.1093/brain/awq075

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 22 May 2019
doi: 10.3389/fnagi.2019.00115

Deep Learning and Multiplex Networks for Accurate Modeling of Brain Age

Nicola Amoroso[1,2], Marianna La Rocca[3], Loredana Bellantuono[1], Domenico Diacono[2], Annarita Fanizzi[4], Eufemia Lella[1,2], Angela Lombardi[2], Tommaso Maggipinto[1,2], Alfonso Monaco[2], Sabina Tangaro[2] and Roberto Bellotti[1,2]*

[1] Dipartimento Interateneo di Fisica "M. Merlin", Università degli studi di Bari "A. Moro", Bari, Italy, [2] Istituto Nazionale di Fisica Nucleare, Bari, Italy, [3] Laboratory of Neuro Imaging, USC Stevens Neuroimaging and Informatics Institute, Keck School of Medicine of USC, University of Southern California, Los Angeles, CA, United States, [4] Istituto Tumori "Giovanni Paolo II" - I.R.C.C.S., Bari, Italy

Recent works have extensively investigated the possibility to predict brain aging from T1-weighted MRI brain scans. The main purposes of these studies are the investigation of subject-specific aging mechanisms and the development of accurate models for age prediction. Deviations between predicted and chronological age are known to occur in several neurodegenerative diseases; as a consequence, reaching higher levels of age prediction accuracy is of paramount importance to develop diagnostic tools. In this work, we propose a novel complex network model for brain based on segmenting T1-weighted MRI scans in rectangular boxes, called patches, and measuring pairwise similarities using Pearson's correlation to define a subject-specific network. We fed a deep neural network with nodal metrics, evaluating both the intensity and the uniformity of connections, to predict subjects' ages. Our model reaches high accuracies which compare favorably with state-of-the-art approaches. We observe that the complex relationships involved in this brain description cannot be accurately modeled with standard machine learning approaches, such as Ridge and Lasso regression, Random Forest, and Support Vector Machines, instead a deep neural network has to be used.

Keywords: age prediction, brain, deep learning, lifespan, aging, structural MRI, machine learning, multiplex networks

OPEN ACCESS

Edited by:
James H. Cole,
King's College London,
United Kingdom

Reviewed by:
Yi Li,
Weill Cornell Medicine, Cornell University, United States
Alessia Sarica,
Universitá degli Studi Magna Graecia, Italy

****Correspondence:***
Nicola Amoroso
nicola.amoroso@uniba.it

Received: *31 January 2019*
Accepted: *01 May 2019*
Published: *22 May 2019*

Citation:
Amoroso N, La Rocca M, Bellantuono L, Diacono D, Fanizzi A, Lella E, Lombardi A, Maggipinto T, Monaco A, Tangaro S and Bellotti R (2019) Deep Learning and Multiplex Networks for Accurate Modeling of Brain Age. Front. Aging Neurosci. 11:115. doi: 10.3389/fnagi.2019.00115

INTRODUCTION

Recently, neuroimaging approaches predicting brain aging have received an increasing attention, especially thanks to the design and development of extremely accurate strategies (Franke et al., 2010; Cole et al., 2017a,b). In fact, the possibility of relying on accurate age predictions allows, as a consequence, the definition of age-related biomarkers for the early detection of anomalous or pathological conditions (Dosenbach et al., 2010; Franke et al., 2012). In particular, machine learning models have been used to learn the aging trajectories of healthy brains thus yielding two main results (Cole and Franke, 2017): (i) predicted age can differ from the actual one and this difference and its entity can suitable define a marker for anomalous/pathological aging (Dukart et al., 2011; Koutsouleris et al., 2013); (ii) subject-specific aging processes can be learned, thus driving personalized monitoring or treatment (when needed) (Baker and Martin, 1997; Cole et al., 2018).

The effectiveness of machine learning methods has resulted to be almost ubiquitous (Hung et al., 2006; Zacharaki et al., 2009; Abraham et al., 2014; Khedher et al., 2015; Al Zoubi et al., 2018). Computer aided detection systems for accurate detection of brain diseases have been

thoroughly investigated, nevertheless there are several studies, for example about Alzheimer's disease, suggesting there is still room for significant improvement (Bron et al., 2015; Amoroso et al., 2018a; Ramírez et al., 2018). More recently, promising results toward these desirable improvements have been found in two distinct directions. On one hand, brain connectivity: describing the brain as a complex network and investigating its properties would enhance the possibility of detection for anomalies and pathological conditions affecting the normal functioning of the brain (Dyrba et al., 2015; Amoroso et al., 2018c); on the other hand, deep learning: the adoption of deep learning techniques, prompted by an increment of both computational resources and observations available to run the learning processes, has become a prominent choice for analyzing medical images for disparate uses, such as segmentation, registration, and classification (Ortiz et al., 2016; Litjens et al., 2017; Shen et al., 2017).

In this work, we present an attempt to combine complex network framework and deep learning strategies to provide a novel accurate modeling of brain age. In particular, we use a multiplex network, which is a multi-layer network. A multiplex is a network with many layers, each of one representing a single subject; the nodes are brain anatomical districts and the connections are their pairwise similarities (Kivelä et al., 2014). Recent studies have demonstrated the advantage of considering multiplex networks instead of single networks in terms of intrinsic information: actually, the information content of the multiplex is not just the sum of the information content of its layers (Battiston et al., 2014; Menichetti et al., 2014).

As for standard networks, multiplex networks can be characterized by suitable metrics (Nicosia and Latora, 2015; Estrada, 2018); in particular, we use nodal properties to obtain a feature representation of a brain and then use this framework to feed a deep learning model to predict the brain age. We compare the performance of deep learning with state-of-the-art regression strategies, such as Lasso regression, Ridge regression, Support Vector Machine, and Random Forest regressions. Besides, we identify the brain regions which seem to majorally affect the age prediction.

MATERIALS AND METHODS

Image Processing

In this work we use data from 5 publicly available sources: ABIDE[1] (Autism Brain Imaging Data Exchange), ADNI[2] (Alzheimer's Disease Neuroimaging Initiative), Beijing Normal University[3], ICBM[4] (International Consortium for Brain Mapping), and IXI[5] (Information eXtraction from Images).

We selected a dataset including 484 subjects in order to obtain a roughly uniform distribution in the age range $7-80$ years; in particular 133 subjects ranged from 7 to 20 years, 120 from 20 to 40 years, 127 from 40 to 60 years, and 104 above 60 years, see **Supplementary Materials** for further details. Subjects whitin the $0-7$ age range are not included in this study because, as better explained in the Discussion section, they require specific image processing techniques which are not require for the age ranges considered here, instead.

Mean age was 37.3 ± 20.4 years. All neuroimaging data used in this study were T1-weighted MPRAGE brain scans (1.5 T or 3.0 T); 1.5 T and 3.0 T scans do not significantly differ in their power to detect gray matter changes (Ho et al., 2010). The participants were healthy controls, thus excluding the presence of neurodegenerative or psychiatric diseases.

Brain scans were normalized in intensity and skull-stripped using the Brain Extraction Tool from the FSL library (Jenkinson et al., 2005); then, non-linear registration was performed using the Advanced Normalization Tools pipeline (Avants et al., 2009) to the MNI152 template; accordingly, all registered scans resulted in $1 \times 1 \times 1\ mm^3$ resolution so that, from now onward, voxels and mm^3 will be interchangeably used.

After spatial normalization we separated the left and the right brain hemispheres and segmented each part in rectangular boxes, called patches, of $l_1 \times l_2 \times l_3$ dimensions. A schematic representation is provided by **Figure 1**.

According to a previous study about neurodegenerative processes in Alzheimer's disease (Amoroso et al., 2018b), we used $l_1 = 10$, $l_2 = 15$, and $l_3 = 20$ with l_1, l_2, and l_3 lengths, in voxels, along the coronal, the axial, and the sagittal orientations, respectively. Thus, each subject's brain was represented by a collection of 600 patches.

The Network Model

By definition, a complex network $G = G(N, L)$ is a couple of two distinct sets (Boccaletti et al., 2006): N, the set of nodes, and L, the set of links. The nodes are the elements of the system one wants to model while the links represent the interactions among them. This basic framework does not take into account the entity of the interactions; to consider this aspect weighted networks are introduced (Newman, 2004). Weighted networks are assigned a third set of elements W whose elements w_{ij}, called weights, represent the strengths of each interaction between the nodes i and j; the weights are usually real or integer numbers, so that a weighted network is denoted $G = G(N, L, W)$.

In this work, the brain networks are defined using each patch as a node. Patches consist of 3,000 voxels whose intensity gray levels ranges from 0 to 1. Accordingly, the whole brain is segmented in 600 patches. We considered each patch as a vector with 3,000 components and measured the Pearson's correlation between each pair of vectors thus obtaining the pairwise similarities, thus we built a weighted network whose nodes were the patches and whose weights were given by the measured correlations. Pearson's correlations range from -1 to 1, however to take into account the left/right symmetry of the brain we kept the absolute value of correlations. Accordingly, our networks consist of 600×600 symmetric adjacency matrices whose rows and columns represent the brain patches and whose elements, ranging from 0 to 1, their absolute Pearson's correlations. It is worth noting that the brain network used in this work is mathematical, in fact nodes have

[1] http://fcon_1000.projects.nitrc.org/indi/abide/
[2] http://adni.loni.usc.edu
[3] http://fcon_1000.projects.nitrc.org/indi/retro/BeijingEnhanced.html
[4] https://ida.loni.usc.edu/
[5] https://brain-development.org/ixi-dataset/

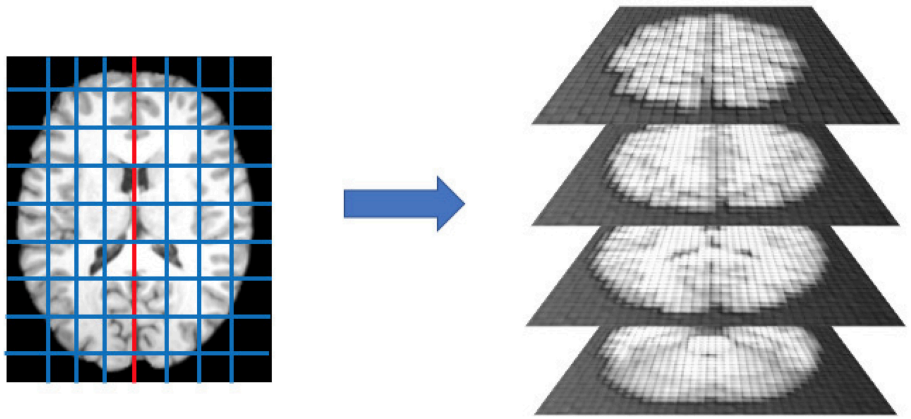

FIGURE 1 | After dividing the brain into left and right hemispheres, each hemisphere is divided in 300 patches. This operation is performed for each subject in the cohort, after registration, thus each patch is expected to roughly contain the same anatomical district and analogous distributions of white matter, gray matter, and cerebrospinal fluid.

no direct anatomical counterparts and edges are correlations, a mathematical similarity metric.

Once the single network representation was obtained for each subject, we built a multiplex model, i.e., a network composed by several layers, in which the same number of nodes can be connected in different ways (Nicosia et al., 2013). Usually, when building a multiplex model, nodes remain unchanged, what changes is the nature of links: for example, in transport networks, the nodes could be the neighbors of a city and the layers the types of transport considered (routes, trains, ...). Age shapes brain networks by modifying the spatial distribution of white matter, gray matter, and cerebrospinal fluid and, therefore, the way brain regions are connected, i.e., their pairwise similarity. Accordingly, it is natural to define a different layer α for each age and, thus, for each subject.

Finally, we measured some specific nodal metrics to characterize the multiplex model. Specifically, we considered the following features:

- Strength s. The sum of the weights associated to the connections of a node is a common centrality metrics used to characterize important nodes within a network. The strength of the node i in a layer α is:

$$s_i^\alpha = \sum_{j=1}^{N} w_{ij}$$

- Inverse Participation Y. It is also important to characterize how strengths are distributed within a network in order to understand the relative importance of a node. The inverse participation of the node i in a layer α is:

$$Y_i^\alpha = \sum_{j=1}^{N} \left(\frac{w_{ij}^\alpha}{s_i^\alpha} \right)^2$$

- Multistrength. The analogous of the strength in a multiplex model.
- Multi-Inverse Participation. The Inverse Participation computed with respect of the multiplex.

Further details, especially about multiplex metrics, are provided for example in Amoroso et al. (2018b). Besides, we computed the conditional probabilities of strength and multistrength against the nodes with degree k; conditional strength for degree k in the layer α is:

$$s(k)^\alpha = \frac{1}{N_k} \sum_{i=1}^{N} s_i^\alpha \delta(k_i^\alpha, k)$$

with N_k the number of nodes with degree k and δ being the Kronecker function, which is equal to one only when the nodal degree k_i^α is k and zero otherwise.

Analogously, the conditional mean of inverse participation for degree k in the layer α is:

$$Y(k)^\alpha = \frac{1}{N_k} \sum_{i=1}^{N} Y_i^\alpha \delta(k_i^\alpha, k)$$

In the end our multiplex representation yielded $M = 8 \times |N|$ features for each subject, with $|N|$ being the cardinality of N, $|N| = 600$, and, therefore, $M = 4,800$. The conceptual workflow is presented in **Figure 2**.

The basic idea behind our approach is that one of the main effects involved by aging is brain atrophy; our framework allows the detection of age related changes in brain using a complex network model and therefore the possibility to yield accurate brain age prediction. Pearson's correlation is a suitable metric to characterize the spatial distribution of white matter, gray matter, and cerebrospinal fluid and the multiplex framework takes into account how this distribution changes over time; besides, the previously mentioned nodal properties measure how these changes affect the networks and the different brain regions, therefore, they allow a direct easy-to-interpret overview of aging effects.

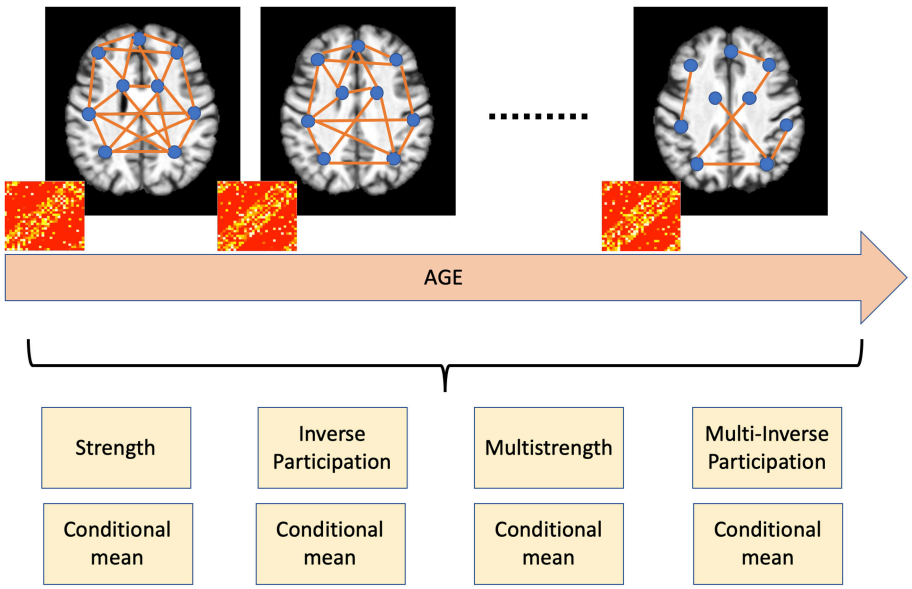

FIGURE 2 | Age shapes brain networks by modifying the spatial distribution of white matter, gray matter, and cerebrospinal fluid. Accordingly, nodal metrics, such as strength and inverse participation, allow the detection and quantification of these age-related changes.

Regression

Once we obtained a feature representation for all subjects, we trained our deep learning regression model. To assess the robustness of our brain model and to confirm the effectiveness of deep learning we also evaluated four other different regression models that are widely adopted for their accuracy: Lasso regression, Ridge regression, Support Vector Machine, and Random Forest. The presented results were cross-validated with a 10-fold procedure repeated 100 times. To evaluate the regression performance we adopted three different metrics:

- Mean Absolute Error (MAE).

$$\mathrm{MAE} = \frac{1}{S}\sum_{i=1}^{S}|y_i - \hat{y}_i|;$$

- Root Mean Squared Error (RMSE).

$$\mathrm{RMSE} = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(y_i - \hat{y}_i)^2};$$

- Pearson's correlation (ρ).

$$\rho = \frac{\sum_{i=1}^{S}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{S}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{S}(\hat{y}_i - \bar{\hat{y}})^2}}.$$

with S being the sample size, y_i the chronological age, $\hat{y}_i$ the predicted brain age, $\bar{y}$ the sample average age, and $\bar{\hat{y}}$ the average brain predicted age. All our models were implemented with the open source R language.

Deep Learning

A deep neural network is, by definition, a network with more than two hidden layers (Hinton et al., 2006). Deep learning strategies are designed to learn, thanks to the complex interactions instanced between neural networks' hidden layers, accurate representations of the provided observations; in recent years, deep learning has significantly improved the state-of-the-art in several fields, such as speech recognition, object detection, and diagnosis support systems (LeCun et al., 2015).

Artificial neural networks with few learning layers, also called shallow networks, have been known for decades; since the introduction of backpropagation algorithms, their training has shown very promising perspectives but raised several feasibility issues, especially for the exponential growth of computational requirements. Besides, a theorem stating that multilayer feed forward networks with a sufficient number of neurons and as few as one hidden layers are universal approximators, strongly suggested to invest more effort on simpler architectures than deeper ones (Hornik et al., 1989). Finally, there was a common belief that deep neural network learning algorithms (especially the gradient descent) could be trapped in local minima preventing the possibility to yield stable and accurate results.

Recent results, both theoretical and empirical, showed that these issues can be overcome and deep learning algorithms can achieve unmatched performances in several domains. Moreover, the possibility to easily access huge computational resources has removed the practical limitations preventing the wide-spread adoption of deep learning strategies.

In this paper, we use a feedforward deep neural networks with four hidden layers respectively including 200, 100, 50, and 20 neurons, see **Figure 3**.

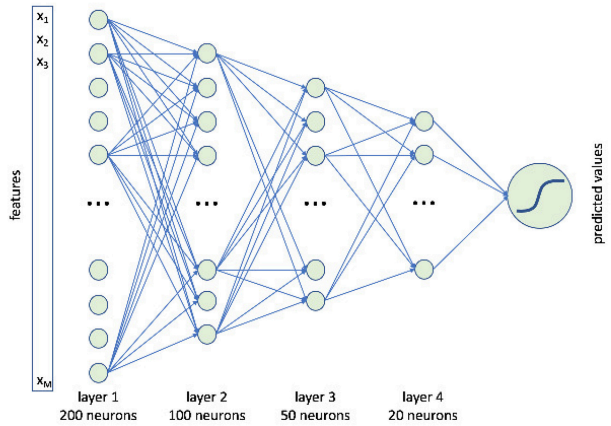

FIGURE 3 | A schematic representation of our deep neural network. It consisted of four hidden layers composed of 200, 100, 50, and 20 neurons.

This architecture was implemented with the "h2o" R package. Among the possible tuning parameters, besides the number of hidden layers with the corresponding neurons, this package offers the possibility to define:

- activation functions including: hyperbolic tangent, linear rectifier, and maxout;
- learning rate;
- training epochs;
- regularization (L_1 or L_2);
- tolerance;
- rate decay.

The flexibility offered by deep learning architectures is also their major drawback, as tuning these models can be challenging. This is why another important option provided by the "h2o" package (and many others) is the so called *grid search*, allowing the systematic exploration of the configurations' space, thus automatically determining the most effective design. We explored different numbers of layers and neurons, as well as different activation functions, while we adopted default values for all the remaining parameters. To increase the network robustness, the weights were randomly initialized at every execution of the algorithm.

We have already mentioned the optimal architecture, for what concerns activation function, hyperbolic tangent was used. We performed extensive search for optimal values thanks to the ReCaS data center[6]; further details about the computational infrastructure are provided in **Supplementary Materials**. Thanks to cross-validation analysis we reached an optimal (and stable) configuration. In order to get a fair comparison with other regression models, we tried to use default configurations whenever possible; parameters whose values were tuned in cross-validation, as for example the number of trees in Random Forests, are explicitly mentioned, otherwise default values must be assumed.

[6]https://www.recas-bari.it/index.php/en/

Ridge Regression

Ridge regression (Hoerl and Kennard, 1970) is a substantial improvement of standard least square regression in those case where independent variables suffer or may suffer from multicollinearity. By definition, multicollinearity consists in the presence of high intercorrelations among the independent variables of the model; when present, multicollinearity can strongly affect the reliability of statistical inferences. Even if brain patches are sufficiently large to mitigate spatial correlations, it is not safe to assume, *a priori*, that neighbor patches are completely independent.

Ridge regression is basically a least square methods. Using the standard notation a regression equation is written in matrix form as $Y = X\beta + e$ with Y the dependent variable, X the independent variables, β the regression coefficients, and e the residuals. Ridge regression prescribes, as standard linear regression, the minimization of the residual sum of squares (RSS):

$$\mathbf{RSS} = \sum_{i=1}^{S} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

where S is the sample size and p the number of independent variables. The difference with standard linear regression is that Ridge regression introduces a penalty or regularization term on the sum of squared coefficients:

$$\mathbf{RSS}_{Ridge} = \sum_{i=1}^{S} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

It is evident that when $\lambda \rightarrow 0$ Ridge regression coincides with ordinary least square regression. When $\lambda \rightarrow \infty$ the high regularization penalty makes some coefficients small, but yet not negligible, thus their effect is limited but still included in the model. Accordingly, the effectiveness of Ridge regression depends on the tuning of λ penalty: models with small λ values tend to have high variance and small bias, on the contrary high λ values involve small variance and high bias. For the present work, we explored several λ values in cross-validation.

Lasso Regression

Ridge regression considers any independent variable from the model whereas Lasso (Least absolute shrinkage and selection operator) regression (Tibshirani, 1996) tackles this issue allowing the exclusions of some coefficients. Accordingly, Lasso regression tries to retain the important features and discard those yielding a negligible contribution to the model.

Lasso residual sum of squares is similar to Ridge regression except for introducing as a penalty contribution the sum of the absolute values of the regression coefficients:

$$\mathbf{RSS}_{Lasso} = \sum_{i=1}^{S} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

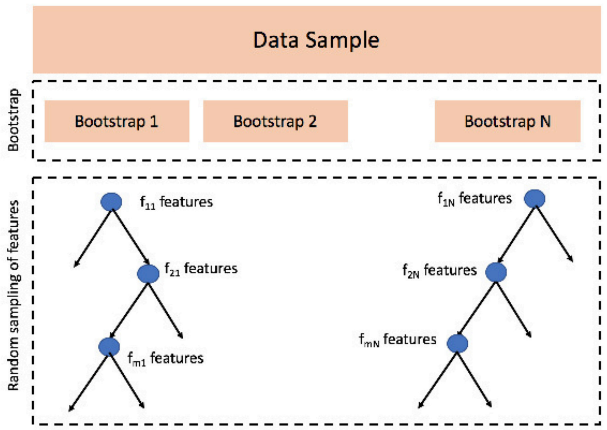

FIGURE 4 | Random Forest algorithm consists of two main phases: (1) data sample is bootstrapped *N* times, *N* is the number of trees and the first parameter this algorithm needs to be set; (2) when growing a tree *j*, at each node splitting *i* a random set of features f_{ij} is sampled; the features selected change at every split, but the number of sampled features, the second parameter to be set, remains constant.

Here, again, *S* is the sample size and *p* the number of independent variables. When $\lambda \rightarrow 0$ Lasso regression coincides with the ordinary least square regression; when $\lambda \rightarrow \infty$ Lasso tends to the null model with all coefficients β_j being 0 and the only non-vanishing value being the intercept. It is worth noting that, as for Ridge regression, Lasso regression needs the tuning of λ parameter in order to balance variance and bias of the model. As for Ridge regression, we explored several λ values in cross-validation.

Random Forest

Another option for regression, extremely popular in recent years, consists in using ensemble learning. Among the possible choices, the most adopted and widely used algorithm is Random Forest (Liaw et al., 2002). Random Forests are constructed bootstrapping the data sample and growing a number of different regression trees, each of them using a different bootstrap, statistically with the original dataset. Besides, as a difference with bagging strategies, Random Forests add a further layer of randomness by growing each tree with a different set of predictors randomly selected every time a node is split, see **Figure 4** for a schematic representation.

The main advantage of Random Forest over classical regression strategies is its robustness on overfitting; moreover, it is a good approach for preliminary investigations in the sense that, depending only two parameters, the number of trees to be grown and the number of features to pick at each node split, Random Forests is easy to tune and control.

A relevant aspect to consider is that Random Forest yields useful information about feature importance, thus resulting in interpretable models and a ranking about the association between each independent variable and the dependent variable, a crucial property in clinical applications. The Random Forest regression was tuned in cross-validation to search optimal values for the number of trees and the number of features to select.

Support Vector Machine

Finally, we evaluated the regression performance using Support Vector Machine (Smola and Schölkopf, 2004). Support Vector Machine regression is based on a well grounded statistical framework whose basic idea consists in using the available observations to learn a function $f(x)$ that has deviations $\epsilon_i < \epsilon$ from targets y_i. As a consequence, the model learns to be accurate at least as the prescribed ϵ precision or, in other words, it does not accept deviations larger than ϵ. For clinical purposes this approach is of fundamental importance, as it guarantees the existence of a limit value which should not be exceeded for the validity of the model.

The main advantages of Support Vector Machine are 2-fold: (i) it is a versatile algorithm which can give accurate results in very different applications, comprising medical ones; (ii) it yields a compact representation even for huge datasets, thus it is a suitable choice for big data applications. The main drawback is probably the need to tune several parameters in order to achieve the perfect balance between variance and bias of the model. A not exhaustive list of parameters to tune include:

- the precision of the model ϵ;
- the kernel used for training and prediction, possible choices are: linear, polynomial (in this case one has to set the degree of the polynomial too), radial basis and sigmoid;
- the cost value for regularization;

Accordingly, for Support Vector Machines to be consistently effective it is fundamental to perform a wide search of the parameter space with a subsequent significant increase of the computational effort. Nevertheless, the use of modern data-centers can easily manage the needed requirements in terms of memory and processing time, thus the computational issues do not discourage the use of this learning framework. We explicitly explore the precision and the cost value for regularization.

RESULTS

Deep Learning Prediction Accuracy

We assessed the performance accuracy of our deep learning model by evaluating three distinct metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Pearson's correlation ρ. The results presented in **Figure 5** show the estimates of these metrics obtained with 100 rounds of 10-fold cross-validation.

Average MAE is 4.7 years, the MAE standard error is 0.1. For what concerns RMSE and correlation, our cross-validated estimates are: RMSE $= 6.2 \pm 1.1$ and $\rho = 0.95 \pm 0.02$.

A not secondary aspect to consider about the reliability of age-predicting models is their homoscedasticity either their heteroscedasticity. We performed the Breush-Pagan test to evaluate the presence or absence of heteroscedasticity and found $p = 0.008$, thus rejecting the null hypothesis, with 5% significance, for the variance of the residuals to be constant over the whole age range.

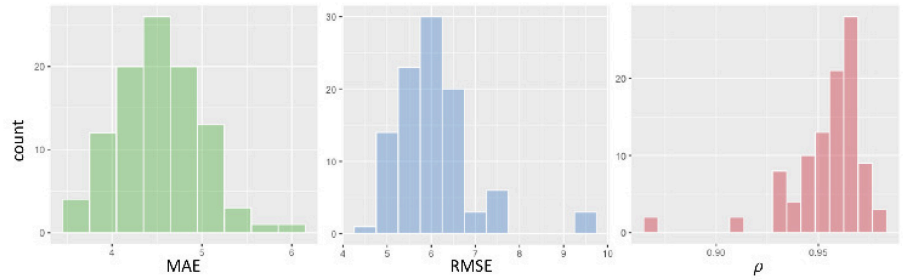

FIGURE 5 | From left to right, histogram of cross-validation results: MAE, RMSE, and Pearson's correlation ρ.

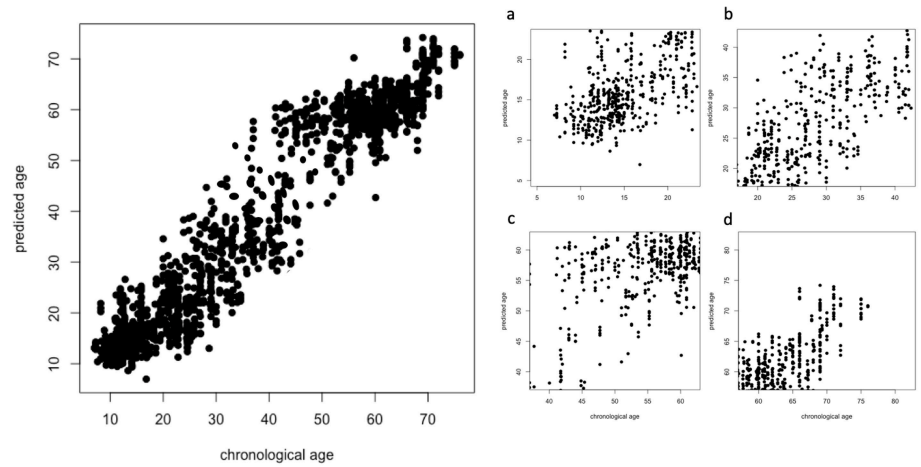

FIGURE 6 | Overall scatter plot of chronological age (x-axis) and predicted age (y-axis) and the specific four age ranges (right panel): $7 \leq \text{Age} < 20$ **(A)**, $20 \leq \text{Age} < 40$ **(B)**, $40 \leq \text{Age} < 60$ **(C)**, $60 \leq \text{Age} < 80$ **(D)**.

Age Ranges Affecting the Model Accuracy

To further investigate the effectiveness of our deep learning model, we evaluated if the regression accuracy was subject to significant changes when considering specific age ranges. In particular, see **Figure 6** for the overall scatter plot (left panel) and four age ranges (right panel): $7 \leq \text{Age} < 20$ (a), $20 \leq \text{Age} < 40$ (b), $40 \leq \text{Age} < 60$ (c), $60 \leq \text{Age} < 80$ (d).

These distributions are significantly different according to a Kruskal-Wallis rank sum test ($p < 2.2e^{-16}$); in particular, the best results are obtained for younger subjects while the performance has a significant drop when considering the groups including older subjects, see **Table 1** for a comprehensive overview.

Correlation is the metric suffering the highest drop in performance over all the considered age ranges. MAE and RMSE share a common behavior, their best values are found when age ranges from 7 to 20; the best correlation is found when $40 \leq \text{Age} < 60$.

TABLE 1 | Performance metrics obtained in different age ranges.

Age range	MAE	RMSE	ρ
7 – 20	**3.7 ± 0.2**	**3.9 ± 0.1**	0.43 ± 0.02
20 – 40	5.1 ± 0.2	6.6 ± 0.1	0.57 ± 0.01
40 – 60	6.5 ± 0.2	8.2 ± 0.2	**0.60 ± 0.01**
60 – 80	4.4 ± 0.2	6.6 ± 0.3	0.41 ± 0.03

In particular, correlation, due to a drastic reduction of the sample size and range, suffers the highest reduction. Best values are in bold.

Sample Size Effect

Previous studies about age prediction using MRI have established the pivotal importance of sample size to obtain accurate age-prediction models. Accordingly, we present in **Figure 7** the assessment of the sample size effect on the accuracy of our model.

The results are 2-fold: performance is affected by sample size, the more the available data, the more accurate age prediction; when using 80% of data, the deep model reaches a robust plateau. Whatever we considered, MAE, RMSE, or ρ correlation, the performance increased with the sample size, besides the variance of the model decreased.

Other Regression Strategies

To demonstrate the pivotal role of deep learning, we used the multiplex features to feed other state-of-the-art regression approaches. In particular, we compared deep learning with

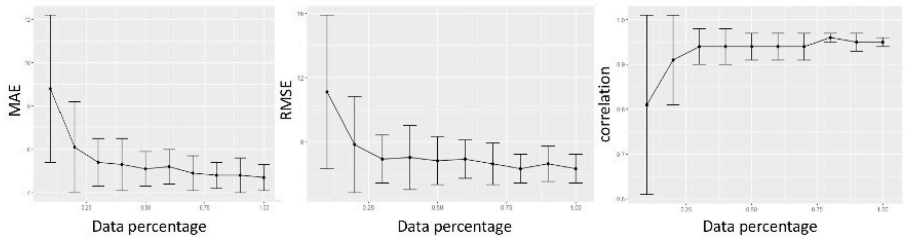

FIGURE 7 | We evaluated the regression metrics MAE, RMSE, and correlation by randomly sampling a varying percentage of subjects from the whole cohort, from 10 to 100%, and reported the results of 100 ten-fold cross-validations.

TABLE 2 | Comparison of cross-validation regression performances for deep learning, Ridge and Lasso regression, Random Forest, and Support Vector Machine.

Model	MAE	RMSE	ρ
Deep learning	**4.7 ± 0.1**	**6.2 ± 1.1**	**0.95 ± 0.02**
Ridge regression	6.0 ± 0.7	7.8 ± 1.3	0.92 ± 0.03
Lasso regression	6.4 ± 0.7	8.2 ± 1.3	0.92 ± 0.03
Random forest	5.9 ± 0.7	7.6 ± 0.9	0.94 ± 0.02
Support vector machine	5.6 ± 0.7	7.2 ± 0.9	0.94 ± 0.01

The reported values are those obtained after grid search for optimal configurations. Best results are presented in bold.

Ridge and Lasso regressions, Random Forest, and Support Vector Machine. **Table 2** shows the comparison among best configurations, further details about parameter tuning and optimal values are reported in **Supplementary Materials**.

Deep learning provides the most accurate model with respect of all the considered metrics. After deep learning, Support Vector Machine gets the best results, nonetheless, deep learning yields a significant increment of about 16% in terms of MAE and 14% in terms of RMSE. For what concerns correlations, even if providing the best performance, deep learning does not seem to significantly improve this metric, another clue suggesting the need for using correlations *cum grano salis*.

Feature Importance and Clinical Validation

To investigate which features had a strategic role in the age prediction, we calculated variable importances by using the Gedeon method (Gedeon, 1997) implemented in the "h2o" R package. This implementation considers the weights connecting the input features to the first two hidden layers and provides, for each features, the relative importance normalized between 0 and 1. We computed the importance ranking over different subject samples in order to select the most strategical features in terms of relative importance and occurrence. We obtained 113 features whose occurrence had not happened by chance (with a 5% comparison threshold with Bonferroni adjustment). In **Table 3**, the first 10 features, directly connected to a patch, are reported in order of mean relative importance along with the corresponding anatomical regions pinpointed by that patch.

The different cortical and sub-cortical anatomical regions, which are proved to be connected with aging, were found by mapping the related patches on the Harvard-Oxford atlas (Desikan et al., 2006). In **Figure 8**, the patches related to these anatomical regions are underlined in red on the MNI 152 template. It is worth to specify that these clinical findings are totally in agreement with the literature as argued in the Discussion section.

DISCUSSION

The method presented in this work, based on the multiplex model combined with a deep learning regression network allows the most accurate age prediction, in comparison with other standard machine learning approaches. Performances presented here compare well with results recently published (Franke et al., 2012; Cole et al., 2017a), including voxel-based approaches, provided the following considerations. First of all, the dataset used in this work is smaller than those investigated in the mentioned works; we have confirmed here that as the sample size increases predicting models tend to be more accurate and with less variance. Nevertheless, as the fraction of data employed exceeds 80%, improvements become significantly smaller; the deep learning model is robust and stable. A not secondary aspect to consider is age distribution: in this work we have analyzed a roughly uniform cohort, which is not the case, e.g., in Cole et al. (2017a). However, the dependence of performance on dataset composition/homogeneity certainly requires further investigation.

Another important aspect to consider about the general validity of the presented results concerns the image processing pipeline. In this study, we used the FSL library; FSL provides a consolidated and widespread tool for brain extraction. Nevertheless, other spatial normalization tools could be used, as for example SPM DARTEL a particularly suitable tool for normalization of elder subjects (Pereira et al., 2010). Actually, there is no general consensus indicating which tool should be preferred, on the contrary it is common for neuroimaging studies to define dedicated pipeline exploiting a wide range of existing tools, such as those previously mentioned, but also including FreeSurfer, ANTs and novel ones (Shen et al., 2013; Im et al., 2015; Hazlett et al., 2017).

In fact, we demonstrated here that age predictions are affected by heteroscedasticity; accordingly, a large data sample uniformly covering the lifespan range could mitigate this

TABLE 3 | First 10 features in order of relative importance for aging prediction along with the related cortical and subcortical brain regions.

Features	Patch	Mean relative importance
Inverse participation	(L) Heschl's Gyrus (includes H1 and H2), Insular Cortex (GM, WM)	0.95
Multistrength	(L) Cingulate Gyrus, anterior division, Cingulate Gyrus, posterior division, Precentral Gyrus (GM)	0.89
Inverse participation	(L) Planum Polare, Heschl's Gyrus (includes H1 and H2), Central Opercular Cortex (GM)	0.89
Multistrength	(L) Frontal Pole, Frontal Orbital Cortex (GM)	0.89
Inverse participation	(R) Paracingulate Gyrus, Cingulate Gyrus, anterior division (GM, WM)	0.89
Strength	(L) Brain Stem, Parahippocampal Gyrus, posterior division (GM)	0.89
Inverse participation	(R) Precentral Gyrus, Post-central Gyrus (GM,WM)	0.88
Inverse participation	(L) Lateral Occipital Cortex, inferior division, Middle Temporal Gyrus, temporo-occipital part (GM, WM)	0.88
Inverse participation	(L) Lateral Occipital Cortex, superior division (GM)	0.88
Inverse participation	(L) Inferior Frontal Gyrus, pars opercularis, Precentral Gyrus, Middle Frontal Gyrus (GM, WM)	0.88

(L) and (R) indicate left and right hemispheres; (GM) and (WM) indicate that gray and white matter are respectively included in the patch corresponding to a certain feature.

issue. Heteroscedasticity also affects performance accuracy: best performances in terms of MAE and RMSE are found for younger subjects (in the $[7 - 20)$ range). This would confirm the necessity to compare age prediction accuracy declared in different studies with the caveat that age distribution of examined cohort should be consistent. This behavior suggests that morphological differences in healthy brains are accentuated in later years, younger brains tend to be less heterogeneous and, therefore, more adherent to a common pattern. However, it is worth noting that the extent of the age-range influences the MAE, with wider age-ranges yielding harder prediction problems; accordingly, we cannot conclude that the model performs better. This consideration about the influence of the age-range on the MAE is also important when comparing the current results between other studies.

Pediatric images usually require specific processing. Actually, children's brains significantly differ from the adult ones, because their growth is characterized by a series of non-linear changes occurring throughout the development ages; this is particularly true between 0 and 7 years. However, we do not expect this effect to significantly affect our analysis, because this specific range was not included in the analysis. Nonetheless, the standard pipeline adopted here is based on a template developed from adult brain data, which are not optimized for pediatric scans and, therefore, this could limit the accuracy of our model. In future work, we plan to focus on age prediction in younger cohorts, limiting the considered age range, and consider dedicated image processing strategies specifically tailored for younger subjects as suggested in recent works (Vân Phan et al., 2018).

A different consideration holds for correlation. Correlations are heavily affected by the overall range of the independent variable, when considering age sub-samples this range decreases, the number of observations decreases too; as a consequence, the resulting correlations do not match with the values computed using the whole dataset. On the other hand, the other metrics take into account only the relative difference between observed and predicted values. In other words, MAE and RMSE on average tend to reproduce in the age subsamples the same behavior they have on the entire dataset. This is not true for correlation. An interesting aspect to investigate in the future could be the assessment of which factors (sample size within each age range, multi-site effect on data heterogeneity, ...) are mostly responsible for this issue. However, deep learning is by far the most accurate method to predict brain age, followed by Support Vector Machine. The intrinsic possibility to manage and model non-linear complex relationships offered by deep models seems to provide a significant advantage when attempting to predict brain age.

Another aspect investigated in this study was the feature importance aimed at finding out which features and which related anatomical regions were more accountable for the age prediction. We chose to not perform a dedicated feature selection in order to outline the role played by the different regression strategies. Of course, feature selection can play an important role in enhancing the performance of machine learning, nevertheless, the focus of this work was to establish the most effective strategy to exploit the informative content provided by our complex network model, independently from other processing steps.

It is interesting to notice as the most important features are often related to patches which identify several times the same anatomical regions demonstrating their prominent role in the aging process. Many studies report that these regions are widely involved in morphometric changes connected with age (Koini et al., 2018). Indeed, significant age-related reduction in cortical thickness, surface area, and volume have been found in areas like Heschl's gyrus, cingulate and paracingulate gyrus, parahippocampal gyrus, and temporal lobe which includes also the planum polare and Heschl's gyrus (Mann et al., 2011; Torii et al., 2012). These two latter regions play an important role in auditory processing which is notoriously

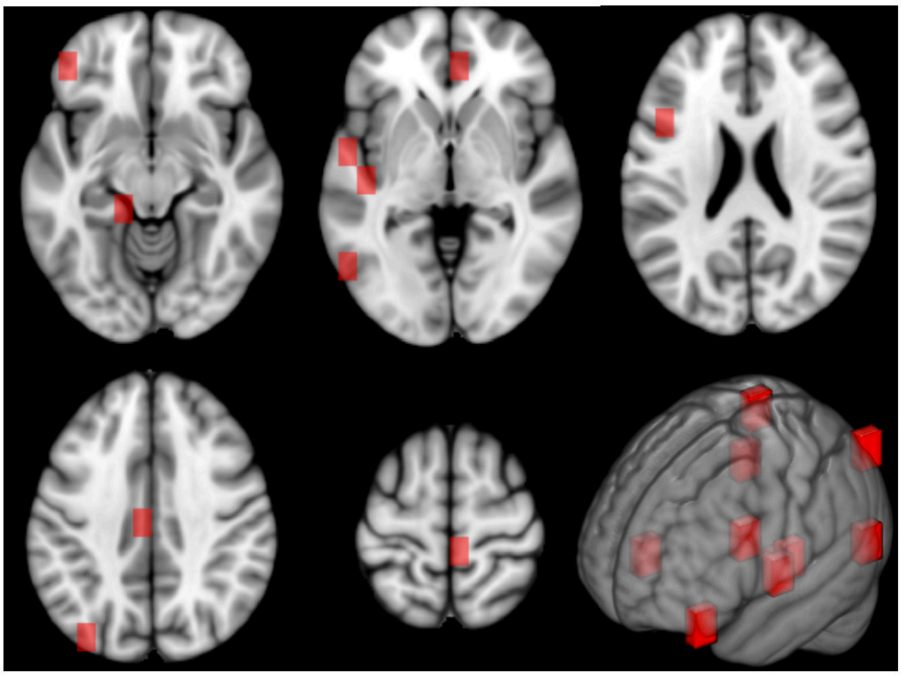

FIGURE 8 | This figure shows the patches related to the first most important features along 5 axial planes of the MNI 152 template. On the bottom right, a 3D representation of the patches on the reference space is reported, as well.

affected by age advancement (Warrier et al., 2009). Cingulate and paracingulate gyrus are implicated in attention and emotional regulation, and parahippocampal gyrus and medial temporal lobe are involved in memory. Therefore, these regions also influence cognitive processes which are still connected with normal aging. A particular vulnerability to cortical thickness changes with age was seen in middle frontal gyrus, pre-central gyrus, post-central gyrus, and in the pars opercularis of the inferior frontal gyrus. The importance of frontal lobe regions is supported by evidence of age-related decline in several cognitive processes such as speed of processing, working memory, cognitive control, and motor control (Thambisetty et al., 2010; Lemaitre et al., 2012). Age-related changes have been also underlined in insula cortical thickness and in brain stem volume (Churchwell and Yurgelun-Todd, 2013; Lambert et al., 2013). However, the reader should take into account that the proposed approach defines a mathematical framework rather than a real biomedical brain network and it should not be overinterpreted

In our results, most of the regions related to the first 10 important features are located in the left hemisphere. This may suggest an age-related decrease or increase of correlation between the patches related to the important features in the left hemisphere and the others. Many studies report that structural and functional hemispheric asymmetry is related to age. Besides, changes in structural brain asymmetry with age have been found right in inferior frontal gyrus, anterior insula, anterior cingulate parahippocampal gyrus, and precentral gyrus (Kovalev et al., 2003), thus, in agreement with our results. Further investigations in this sense could be interesting also to examine a still open issue: whether and which hemisphere ages faster that currently is still an open issue (Esteves et al., 2018). However, the reader should take into account the proposed approach defines a mathematical framework rather than a real biomedical brain network and it should not be overinterpreted.

Finally, it is worth to mention an aspect that is gaining more and more interest, which is the increasingly widespread of "artificial intelligence" and machine learning for health purposes, especially for the development of diagnosis support systems. On one hand, thanks to deep learning there is the possibility to use raw data to directly predict age, height, or subject-specific clinical scores, the presence of pathological conditions and eventually their severity. On the other hand, thanks to particular inversion strategies, recent works have demonstrated the possibility to retrieve sensible information on patients even when using pre-trained models (Fredrikson et al., 2015). With this perspective, using our multiplex model, mediating between raw data and clinical score, in this case age prediction, could be also considered a safe way to use sensible data and protect the users' privacy, not to mention the computational advantage in terms of processing time.

CONCLUSIONS

In this work, we demonstrated that: (i) the features retrieved with our novel brain network model can accurately characterize the normal aging, besides their informative content compares well with state-of-the-art; (ii) the informative power of multiplex features is effectively exploited and significantly maximized when using a deep learning regression. The proposed methodology

localizes the brain regions most affecting aging in the left hemisphere. For what concerns the model accuracy, further investigations should be performed by increasing the sample size; the presented results are promising, nevertheless the statistical robustness of this study would greatly benefit from a larger dataset, besides this would be of paramount importance for a fair comparison with other studies. Finally, we observed here that brain aging is strongly affected by heteroscedasticity, this effect should properly taken into account by studies investigating lifespan processes; in particular, worst prediction accuracy was obtained in the age range $40 - 60$, this would reflect the high specificity and variability characterizing brain atrophy in these years. Nevertheless, further investigations, exceeding the aims of the present work will be needed to corroborate such hypothesis.

ETHICS STATEMENT

All experiments were performed with the informed consent of each participant or caregiver in line with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Local institutional ethics committees approved the study.

AUTHOR CONTRIBUTIONS

NA designed the study and conceived the model. NA and ML performed statistical analyses. All authors interpreted the data, wrote, and approved the manuscript.

ACKNOWLEDGMENTS

In this study, the data included subjects recruited in: Alzheimer's Disease Neuroimaging Initiative (ADNI) with T1 images (http://adni.loni.usc.edu).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2019.00115/full#supplementary-material

REFERENCES

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., et al. (2014). Machine learning for neuroimaging with scikit-learn. *Front. Neuroinformatics* 8:14. doi: 10.3389/fninf.2014.00014

Al Zoubi, O., Ki Wong, C., Kuplicki, R. T., Yeh, H. w., Mayeli, A., Refai, H., et al. (2018). Predicting age from brain EEG signals–a machine learning approach. *Front. Aging Neurosci.* 10:184. doi: 10.3389/fnagi.2018.00184

Amoroso, N., Diacono, D., Fanizzi, A., La Rocca, M., Monaco, A., Lombardi, A., et al. (2018a). Deep learning reveals Alzheimer's disease onset in MCI subjects: results from an international challenge. *J. Neurosci. Methods* 302, 3–9. doi: 10.1016/j.jneumeth.2017.12.011

Amoroso, N., La Rocca, M., Bruno, S., Maggipinto, T., Monaco, A., Bellotti, R., et al. (2018b). Multiplex networks for early diagnosis of Alzheimer's disease. *Front. Aging Neurosci.* 10:365. doi: 10.3389/fnagi.2018.00365

Amoroso, N., La Rocca, M., Monaco, A., Bellotti, R., and Tangaro, S. (2018c). Complex networks reveal early MRI markers of Parkinson's disease. *Med. Image Anal.* 48, 12–24. doi: 10.1016/j.media.2018.05.004

Avants, B. B., Tustison, N., and Song, G. (2009). Advanced normalization tools (ANTS). *Insight J* 2, 1–35.

Baker, G. T., and Martin, G. R. (1997). "Molecular and biologic factors in aging: the origins, causes, and prevention of senescence," in *Geriatric Medicine* (New York, NY:Springer), 3–28.

Battiston, F., Nicosia, V., and Latora, V. (2014). Structural measures for multiplex networks. *Phys. Rev. E* 89:032804. doi: 10.1103/PhysRevE.89.032804

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: structure and dynamics. *Phys. Rep.* 424, 175–308. doi: 10.1016/j.physrep.2005.10.009

Bron, E. E., Smits, M., Van Der Flier, W. M., Vrenken, H., Barkhof, F., Scheltens, P., et al. (2015). Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. *Neuroimage* 111, 562–579. doi: 10.1016/j.neuroimage.2015.01.048

Churchwell, J. C., and Yurgelun-Todd, D. A. (2013). Age-related changes in insula cortical thickness and impulsivity: significance for emotional development and decision-making. *Dev. Cogn. Neurosci.* 6, 80–86. doi: 10.1016/j.dcn.2013.07.001

Cole, J. H., and Franke, K. (2017). Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends Neurosci.* 40, 681–690. doi: 10.1016/j.tins.2017.10.001

Cole, J. H., Marioni, R. E., Harris, S. E., and Deary, I. J. (2018). Brain age and other bodily 'ages': implications for neuropsychiatry. *Mol. Psychiatry* 24, 266–281. doi: 10.1038/s41380-018-0098-1

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

Cole, J. H., Ritchie, S. J., Bastin, M. E., Hernández, M. V., Maniega, S. M., Royle, N., et al. (2017b). Brain age predicts mortality. *Mol. Psychiatry* 23, 1385–1392. doi: 10.1038/mp.2017.62

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

Dosenbach, N. U., Nardos, B., Cohen, A. L., Fair, D. A., Power, J. D., Church, J. A., et al. (2010). Prediction of individual brain maturity using fMRI. *Science* 329, 1358–1361. doi: 10.1126/science.1194144

Dukart, J., Schroeter, M. L., Mueller, K., and Alzheimer's Disease Neuroimaging Initiative (2011). Age correction in dementia–matching to a healthy brain. *PLoS ONE* 6:e22193. doi: 10.1371/journal.pone.0022193

Dyrba, M., Grothe, M., Kirste, T., and Teipel, S. J. (2015). Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum. Brain Mapp.* 36, 2118–2131. doi: 10.1002/hbm.22759

Esteves, M., Magalhães, R., Marques, P., Castanho, T. C., Portugal-Nunes, C., Soares, J. M., et al. (2018). Functional hemispheric (a) symmetries in the aged brain–relevance for working memory. *Front. Aging Neurosci.* 10:58. doi: 10.3389/fnagi.2018.00058

Estrada, E. (2018). Communicability geometry of multiplexes. *New J. Phys.* 21:015004. doi: 10.1088/1367-2630/aaf8bc

Franke, K., Luders, E., May, A., Wilke, M., and Gaser, C. (2012). Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. *Neuroimage* 63, 1305–1312. doi: 10.1016/j.neuroimage.2012.08.001

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., and Alzheimer's Disease Neuroimaging Initiative (2010). Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *Neuroimage* 50, 883–892. doi: 10.1016/j.neuroimage.2010.01.005

Fredrikson, M., Jha, S., and Ristenpart, T. (2015). "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (New York, NY:ACM), 1322–1333.

Gedeon, T. D. (1997). Data mining of inputs: analysing magnitude and functional measures. *Int. J. Neural Syst.* 8, 209–218. doi: 10.1142/S0129065797000227

Hazlett, H. C., Gu, H., Munsell, B. C., Kim, S. H., Styner, M., Wolff, J. J., et al. (2017). Early brain development in infants at high risk

for autism spectrum disorder. *Nature* 542:348. doi: 10.1038/nature 21369

Hinton, G. E., Osindero, S., and Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Ho, A. J., Hua, X., Lee, S., Leow, A. D., Yanovsky, I., Gutman, B., et al. (2010). Comparing 3 T and 1.5 T MRI for tracking Alzheimer's disease progression with tensor-based morphometry. *Hum. Brain Mapp.* 31, 499–514. doi: 10.1002/hbm.20882

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 359–366. doi: 10.1016/0893-6080(89)90020-8

Hung, W.-L., Yang, M.-S., and Chen, D.-H. (2006). Parameter selection for suppressed fuzzy c-means with an application to MRI segmentation. *Pattern Recogn. Lett.* 27, 424–438. doi: 10.1016/j.patrec.2005.09.005

Im, K., Raschle, N. M., Smith, S. A., Ellen Grant, P., and Gaab, N. (2015). Atypical sulcal pattern in children with developmental dyslexia and at-risk kindergarteners. *Cereb. Cortex* 26, 1138–1148. doi: 10.1093/cercor/bhu305

Jenkinson, M., Pechaud, M., Smith, S., et al. (2005). "Bet2: Mr-based estimation of brain, skull and scalp surfaces," in *Eleventh Annual Meeting of the Organization for Human Brain Mapping*, Vol. 17 (Toronto, ON), 167.

Khedher, L., Ramírez, J., Górriz, J. M., Brahim, A., Segovia, F., and the Alzheimer's Disease Neuroimaging Initiative (2015). Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing* 151, 139–150. doi: 10.1016/j.neucom.2014.09.072

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *J. Complex Netw.* 2, 203–271. doi: 10.1093/comnet/cnu016

Koini, M., Duering, M., Gesierich, B. G., Rombouts, S. A., Ropele, S., Wagner, F., et al. (2018). Grey-matter network disintegration as predictor of cognitive and motor function with aging. *Brain Struct. Funct.* 223, 2475–2487. doi: 10.1007/s00429-018-1642-0

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., et al. (2013). Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophr. Bull.* 40, 1140–1153. doi: 10.1093/schbul/sbt142

Kovalev, V. A., Kruggel, F., and von Cramon, D. Y. (2003). Gender and age effects in structural brain asymmetry as measured by mri texture analysis. *Neuroimage* 19, 895–905. doi: 10.1016/S1053-8119(03)00140-X

Lambert, C., Chowdhury, R., Fitzgerald, T., Fleming, S. M., Lutti, A., Hutton, C., Draganski, B., Frackowiak, R., and Ashburner, J. (2013). Characterizing aging in the human brainstem using quantitative multimodal mri analysis. *Front. Hum. Neurosci.* 7:462. doi: 10.3389/fnhum.2013.00462

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

Lemaitre, H., Goldman, A. L., Sambataro, F., Verchinski, B. A., Meyer-Lindenberg, A., Weinberger, D. R., et al. (2012). Normal age-related brain morphometric changes: nonuniformity across cortical thickness, surface area and gray matter volume? *Neurobiol. Aging* 33, 617.e1–e9. doi: 10.1016/j.neurobiolaging.2010.07.013

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomForest. *R News* 2, 18–22.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Mann, S. L., Hazlett, E. A., Byne, W., Hof, P. R., Buchsbaum, M. S., Cohen, B. H., et al. (2011). Anterior and posterior cingulate cortex volume in healthy adults: effects of aging and gender differences. *Brain Res.* 1401, 18–29. doi: 10.1016/j.brainres.2011.05.050

Menichetti, G., Remondini, D., Panzarasa, P., Mondragón, R. J., and Bianconi, G. (2014). Weighted multiplex networks. *PLoS ONE* 9:e97857. doi: 10.1371/journal.pone.0097857

Newman, M. E. (2004). Analysis of weighted networks. *Phys. Rev. E* 70:056131. doi: 10.1103/PhysRevE.70.056131

Nicosia, V., Bianconi, G., Latora, V., and Barthelemy, M. (2013). Growing multiplex networks. *Phys. Rev. Lett.* 111:058701. doi: 10.1103/PhysRevLett.111.058701

Nicosia, V., and Latora, V. (2015). Measuring and modeling correlations in multiplex networks. *Phys. Rev. E* 92:032805. doi: 10.1103/PhysRevE.92.032805

Ortiz, A., Munilla, J., Gorriz, J. M., and Ramirez, J. (2016). Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. *Int. J. Neural Syst.* 26:1650025. doi: 10.1142/S0129065716500258

Pereira, J. M., Xiong, L., Acosta-Cabronero, J., Pengas, G., Williams, G. B., and Nestor, P. J. (2010). Registration accuracy for VBM studies varies according to region and degenerative disease grouping. *Neuroimage* 49, 2205–2215. doi: 10.1016/j.neuroimage.2009.10.068

Ramírez, J., Górriz, J., Ortiz, A., Martínez-Murcia, F., Segovia, F., Salas-Gonzalez, D., et al. (2018). Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. *J. Neurosci. Methods* 302, 47–57. doi: 10.1016/j.jneumeth.2017.12.005

Shen, D., Wu, G., and Suk, H. I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248. doi: 10.1146/annurev-bioeng-071516-044442

Shen, M. D., Nordahl, C. W., Young, G. S., Wootton-Gorges, S. L., Lee, A., Liston, S. E., et al. (2013). Early brain enlargement and elevated extra-axial fluid in infants who develop autism spectrum disorder. *Brain* 136, 2825–2835. doi: 10.1093/brain/awt166

Smola, A. J., and Schölkopf, B. (2004). A tutorial on support vector regression. *Stat. Comput.* 14, 199–222. doi: 10.1023/B:STCO.0000035301.49549.88

Thambisetty, M., Wan, J., Carass, A., An, Y., Prince, J. L., and Resnick, S. M. (2010). Longitudinal changes in cortical thickness associated with normal aging. *Neuroimage* 52, 1215–1223. doi: 10.1016/j.neuroimage.2010.04.258

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Torii, Y., Iritani, S., Sekiguchi, H., Habuchi, C., Hagikura, M., Arai, T., et al. (2012). Effects of aging on the morphologies of heschl's gyrus and the superior temporal gyrus in schizophrenia: a postmortem study. *Schizophr. Res.* 134, 137–142. doi: 10.1016/j.schres.2011.10.024

Vân Phan, T., Smeets, D., Talcott, J. B., and Vandermosten, M. (2018). Processing of structural neuroimaging data in young children: bridging the gap between current practice and state-of-the-art methods. *Dev. Cogn. Neurosci.* 33, 206–223. doi: 10.1016/j.dcn.2017.08.009

Warrier, C., Wong, P., Penhune, V., Zatorre, R., Parrish, T., Abrams, D., and Kraus, N. (2009). Relating structure to function: Heschl's gyrus and acoustic processing. *J. Neurosci.* 29, 61–69. doi: 10.1523/JNEUROSCI.3489-08.2009

Zacharaki, E. I., Wang, S., Chawla, S., Soo Yoo, D., Wolf, R., Melhem, E. R., et al. (2009). Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magn. Reson. Med.* 62, 1609–1618. doi: 10.1002/mrm.22147

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 29 May 2019
doi: 10.3389/fninf.2019.00033

Evaluation of Enhanced Learning Techniques for Segmenting Ischaemic Stroke Lesions in Brain Magnetic Resonance Perfusion Images Using a Convolutional Neural Network Scheme

Carlos Uziel Pérez Malla[1], *Maria del C. Valdés Hernández*[2]*, *Muhammad Febrian Rachmadi*[1] and *Taku Komura*[1]

[1] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [2] Department of Neuroimaging Sciences, University of Edinburgh, Edinburgh, United Kingdom

Magnetic resonance (MR) perfusion imaging non-invasively measures cerebral perfusion, which describes the blood's passage through the brain's vascular network. Therefore, it is widely used to assess cerebral ischaemia. Convolutional Neural Networks (CNN) constitute the state-of-the-art method in automatic pattern recognition and hence, in segmentation tasks. But none of the CNN architectures developed to date have achieved high accuracy when segmenting ischaemic stroke lesions, being the main reasons their heterogeneity in location, shape, size, image intensity and texture, especially in this imaging modality. We use a freely available CNN framework, developed for MR imaging lesion segmentation, as core algorithm to evaluate the impact of enhanced machine learning techniques, namely data augmentation, transfer learning and post-processing, in the segmentation of stroke lesions using the ISLES 2017 dataset, which contains expert annotated diffusion-weighted perfusion and diffusion brain MRI of 43 stroke patients. Of all the techniques evaluated, data augmentation with binary closing achieved the best results, improving the mean Dice score in 17% over the baseline model. Consistent with previous works, better performance was obtained in the presence of large lesions.

Keywords: ischaemic stroke, medical image analysis, deep learning, computer vision, convolutional neural networks, deepmedic, segmentation

OPEN ACCESS

Edited by:
Juan Manuel Gorriz,
University of Granada, Spain

Reviewed by:
Islem Rekik,
University of Dundee, United Kingdom
Maneul Grana,
Universidad del Pais Vasco, Spain

***Correspondence:**
Maria del C. Valdés Hernández
m.valdes-hernan@ed.ac.uk

Received: *01 February 2019*
Accepted: *23 April 2019*
Published: *29 May 2019*

Citation:
Pérez Malla CU, Valdés Hernández MC, Rachmadi MF and Komura T (2019) Evaluation of Enhanced Learning Techniques for Segmenting Ischaemic Stroke Lesions in Brain Magnetic Resonance Perfusion Images Using a Convolutional Neural Network Scheme. Front. Neuroinform. 13:33. doi: 10.3389/fninf.2019.00033

1. INTRODUCTION

Magnetic resonance imaging (MRI) has become a powerful clinical tool for diagnostics. Its application has been expanded to the evaluation of brain function through the assessment of a number of functional and metabolic parameters. One such parameter is cerebral perfusion, which describes the passage of blood through the brain's vascular network. Amongst the several techniques used to measure cerebral perfusion (Petrella and Provenzale, 2000; Fantini et al., 2016), MRI is perhaps the most widely used due to its non-invasiveness. Thus, having great potential in becoming an important tool in the diagnosis and treatment of patients with cerebrovascular disease and other brain disorders. It measures cerebral perfusion via assessment of various hemodynamic

measurements such as cerebral blood volume, cerebral blood flow, and mean transit time, from serial tissue tracer concentration measurements. These measurements are analyzed in relation to their values in normal tissue regions (e.g., normal-appearing white matter). Therefore, the importance of estimating the location and extent of the abnormal region automatically.

Expert delineation is usually performed in the imaging modality that best displays the pathology while simultaneously evaluating other imaging modalities. The quality of this process depends on the expert's experience, and suffers from intra- and inter-observer variability (Kamnitsas et al., 2017). Automated segmentation methods are not only necessary to provide the quantitative information needed to better support clinical decisions, but also to carry out large scale studies, with increased reliability and reproducibility, for which manual delineation is simply unattainable (Maier et al., 2017). Most of these algorithms use expert-labeled data to "learn" the pattern to be segmented until a certain level of accuracy is reached, and are expected to reproduce similar accuracy levels for new unlabeled data. Deep Learning algorithms, such as Convolutional Neural Networks (CNN), have risen in popularity due to their success on computer vision research (Krizhevsky et al., 2012). Though CNNs are typically used for multi-label image classification problems, they can also be employed for segmentation tasks by classifying each voxel according to the region they belong to Kamnitsas et al. (2017).

In MR perfusion imaging, the pathologies' appearance does not follow a clear pattern, which makes their detection far more difficult. Specifically ischaemic lesions can appear anywhere in the brain and their shape and signal intensities vary not only between disease stages but also within them (Maier et al., 2017). This variability increases with time from the stroke onset. Also, the intensity within the infarcted region is not necessarily homogeneous (Kamnitsas et al., 2017).

1.1. CNN Architectures for Brain Lesion Segmentation - DeepMedic

Specifically for the segmentation of brain lesions, different CNNs architectures have been evaluated (He et al., 2016; López-Zorrilla et al., 2017; Guerrero et al., 2018). One of them (Guerrero et al., 2018) proposed a 2D CNN architecture for White Matter Hyperintensities (WMH) segmentation, and reported having achieved state of the art performance in differentiating them from ischaemic stroke lesions. However, by taking a 2D approach, it discards important spatial information, since did not take into account the volumetric nature of the data; and was only evaluated using structural MRI modalities, where lesions are homogeneous and easier to identify.

Using a 3D approach to manipulate Magnetic Resonance Imaging (MRI) data is not straightforward, as it requires significantly more computing power and memory than the 2D counterparts (Roth et al., 2014). The main factor that attempts against 3D segmentation is the slow inference process. This can be alleviated by taking advantage of dense inference (Sermanet et al., 2013), a property of full convolutional networks that avoids recomputing convolutions for overlapping image patches and thus reduces inference times. 3D CNN architectures have been used to segment pathologies (Brosch et al., 2016; Milletari et al., 2016). However, DeepMedic (Kamnitsas et al., 2017) has emerged as the brain lesion segmentation CNN method for excellence, due to its availability, technical support and versatility, as it has been applied not only to segment hyperintense lesions (Rachmadi et al., 2018b), but also lesions with heterogeneous signal intensities (i.e., tumors) (Kamnitsas et al., 2017). It has a 3D CNN architecture of two pathways that uses dense-inference and adds a 3D fully connected Conditional Random Forest (CRF) as a final post-processing layer. By taking advantage of the dense inference, DeepMedic can be trained using image segments (i.e., image patches of size bigger than the network's receptive field) to avoid recomputing convolutions of overlapping patches. Additionally, the dual pathway is used to compute both local and global (i.e., contextual) features at the same time by processing the same image at different scales. Finally, the CRF is used to remove false positives before returning the final results. DeepMedic reached the first position in the **I**schemic **S**troke lesion **S**egmentation (SISS) subchallenge of the **I**schemic **S**troke **LE**sion **S**egmentation (ISLES) 2015 challenge[1].

In subsequent ISLES challenges other CNN approaches have been applied. For example, whilst DeepMedic uses a traditional cross-entropy function (Kamnitsas et al., 2017), the winners of the ISLES 2017 challenge (Choi et al., 2017; Lucas and Heinrich, 2017), use a loss function based on Dice Similarity Coefficient (DSC) particularly designed for unbalanced data sets (Sudre et al., 2017). Also, (Choi et al., 2017) implement a spatial pyramid pooling layer (He et al., 2014), recently combined with an encoder-decoder (Chen et al., 2018b) to improve segmentation predictions. Spatial pyramid pooling guarantees a fixed output size for different sized inputs (He et al., 2014). This means that the network can process inputs at different scales, similarly to DeepMedic, while keeping the same output size. Dilated convolutions have also proven useful for enhancing the spatial resolution of the network and thus improving the performance for semantic segmentation (Chen et al., 2017, 2018a). These convolutional layers extend the field of view and thus can extract features at different scales.

1.2. Enhancing Learning Techniques

Variations in CNN architectures appear to show improvements in the segmentation of certain pathologies. However, these methods suffer a significant loss in performance when these changes are applied to datasets acquired with different imaging protocols, or using different sequences (i.e., task domain changes), they are applied to the assessment of different types of lesions caused by different pathology (e.g., the initial task being to segment tumor lesions, whilst the actual task is to segment ischaemic stroke lesions), or they are expected to perform tasks that are related to but not the same task they were trained for (e.g., lesion segmentation vs. lesion assessment).

There are several ways to enhance the performance of the CNN architectures without modifying the architecture

[1] www.isles-challenge.org/ISLES2015/

itself. In general, they can be enumerated as follows: (1) pre-processing the input data, (2) modifying the input data by adding information derived from internal and external sources (i.e., data augmentation), (3) re-purposing a model trained for one task to perform a second related task (i.e., transfer learning), and (4) post-processing the output from the CNN.

1.2.1. Pre-processing the Input Data

The importance of pre-processing the data has been highlighted by previous works. For example, Rachmadi et al. (2018b), for segmenting WMH, extract the brain tissue from the originally acquired MRI, and only input this to the CNN architecture. In addition, perform a three-step intensity normalization: (1) adjust the maximum gray scale value of the MRI brain to 10 percent of the maximum intensity value, (2) adjust the contrast and brightness of the images such that their histograms are consistent, and (3) normalize the intensities of the resultant images to zero-mean and unit-variance. Guerrero and colleagues, for similar task, used two MRI modalities (Guerrero et al., 2018), which were co-registered, resliced to have 1×1 mm in-plane voxel size, and normalized their intensities. In general, intensity normalization, contrast adjustment and removal of background features that could confound the algorithms are necessary for achieving a good segmentation. When multiple MRI sequences or imaging modalities are used, co-registration is also necessary.

1.2.2. Data Augmentation

Training a machine learning model is equivalent to tune its parameters so that it can map a particular input to an output. The number of parameters needed is proportional to the complexity of the task. These parameters can increase if more information is given. The increase in the amount of input data without necessarily meaning an increase in the contextual or semantic data *per se* is known as data augmentation and has been used in brain image segmentation tasks. Several studies have introduced global spatial information as an additional input to CNN schemes in form of large 2D orthogonal patches down-scaled by a certain factor (de Brebisson and Montana, 2015), integrated with intensity features from image voxels (Van Nguyen et al., 2015), as a number of hand-crafted spatial location features (Ghafoorian et al., 2016), synthetic volume (Steenwijk et al., 2013; Roy et al., 2015), or set of synthetic images that encode spatial information (Rachmadi et al., 2018b) for mentioning some examples. In other words, all input datasets are acquired under a limited set of conditions (e.g., specific MRI scanning protocols, pathology appearance restricted to few examples, etc.). However, our target application may exist in a variety of conditions (e.g., pathologies in different location, scale, brightness, contrasts, shapes). By synthetically generating data to account for these variations without adding irrelevant features, good results might be obtained. A review of the state of the art in medical image analysis concluded that very similar algorithms could achieve different results due to smart data pre-processing and augmentation (Litjens et al., 2017).

1.2.3. Transfer Learning

Transfer learning has become a popular choice for re-purposing machine learning models that have proven useful for particular tasks, by means of either fine-tuning pre-trained models with data of another nature (i.e., domain adaptation transfer learning), or using a pre-trained model as a starting point for a model on a second task of interest (i.e., task adaptation transfer learning). Domain adaptation transfer learning, where data domains in training and testing processes differ, has been applied successfully to brain MRI segmentation tasks. For example, one study improved Support Vector Machines (SVM)'s performance using different distribution of training data (Van Opbroek et al., 2015). Another study pre-trained CNN using natural images for segmentation of neonatal to adult brain images (Xu et al., 2017), and other study pre-trained a CNN for brain lesion segmentation using MRI data acquired with other protocols (Ghafoorian et al., 2017). Task adaptation transfer learning has been applied to WMH segmentation, by teaching a CNN to "learn" to detect texture irregularities instead of binary expert-delineated WMH segmentations (Rachmadi et al., 2018a).

1.3. Contributions

Our main contributions are to propose and evaluate data augmentation and transfer learning methods for improving the output of a widely used brain lesion segmentation CNN approach, namely DeepMedic, to identify and delineate the ischaemic stroke lesion from MR perfusion imaging.

2. METHODS

2.1. Data

The ISLES challenge was conceived as a common benchmark for researchers to compare their segmentation algorithms (Maier et al., 2017) for ischaemic stroke lesions. Initially, the first iteration of ISLES (in 2015), included two sub-challenges, namely **S**troke **P**erfusion **ES**timation (SPES) and SISS. The first sub-challenge was about segmenting stroke lesions in the acute phase, whereas the second focused on sub-acute lesions (Maier et al., 2017).

The stroke cases were carefully crafted and included a wide range of lesion variability. Images were obtained in clinical routine, with different amounts of image artifacts and different views (Maier et al., 2017). Also, some subjects suffered from other pathologies that could be mistaken for ischemic stroke lesions. All files are given in uncompressed Neuroimaging Informatics Technology Initiative (NIfTI) format: (*.nii).

ISLES 2017 contains 43 and 32 training and testing acute subjects, respectively. Included MRI sequences are Apparent Diffusion Coefficient (ADC), 4D Perfusion Weighted Image (4DPWI), Mean Transient Time (MTT), relative Cerebral Blood Flow (rCBF), relative Cerebral Blood Volume (rCBV), Time to maximum (Tmax) and Time to peak (TTP). Images from all modalities were skull-stripped, anonymized and individually co-registered.

The Ground Truth (GT) files, which delimit the actual lesion region, were only provided for training subjects, so as to avoid having participants performing fine-tuning on the test data. They

were segmented on T2-weighted and Fluid Attenuation Inversion Recovery (FLAIR) sequences after the stroke had stabilized, but these imaging modalities were not provided.

After careful examination, the stroke subjects in the training data were classified into three different stroke subtypes. These are lacunar/subcortical (10 subjects), small cortical (7 subjects) and big cortical/main artery (26 subjects).

2.2. Baseline Configuration

The baseline CNN model, including its architecture and hyper-parameters, is based on DeepMedic v0.6.1 (Kamnitsas et al., 2017). The architecture used slightly differs from the initial architecture (Kamnitsas et al., 2017).

The number of convolutional layers was 8, and the number of feature maps for each were [30, 30, 40, 40, 40, 50, 50]. The kernel size was (3, 3, 3) for all layers. Residual connections in both pathways were also included so that the input of layers [3, 4, 6] was added to the output of layers [4, 6, 8].

The final blocks of the scheme were composed of Fully Connected (FC) layers and a CRF. The number of FC layers was set to two, with 150 feature maps each. The size of the kernels of the first FC layer, which combined the outputs of different scales, was again (3, 3, 3). Additionally, there was a residual connection between the second and first layers, meaning that the input of the first FC layer was added to the output of the second and final FC layer.

The second pathway had an additional parameter that determined the downsampling factor applied to the images fed to the second pathway. Additionally, batch normalization (Ioffe and Szegedy, 2015) was added at the end of each convolutional layer.

The dimension of the training and validation segments were [25, 25, 25] and [17, 17, 17], respectively. The latter was equal to the receptive field of the network. The size of the segments was limited by the available RAM and GPU memory.

The batch size for training, validation and inference were set to 24, 48, and 24, respectively. Dropout (Srivastava et al., 2014) was added in the second FC layer and the final classification layer, both with a rate of 0.5. Weight initialization followed a modified Xavier initialization (Glorot and Bengio, 2010) that accounts for nonlinearities (He et al., 2015). This allows the training of deeper networks and works well with Parametric Rectified Linear Units (PReLU) (He et al., 2015), which were the predefined activation units.

Also, intracranial volume masks were provided to limit the region where samples were extracted from, which in turn saved time and memory. This means that foreground samples were extracted from the GT label mask and background samples extracted from the region inside the subject mask minus the intersection with the label mask. By default, samples were extracted centered in a foreground or background voxel with equal probability.

During training, epochs were divided into subepochs. The number of epochs and subepochs was set to 35 and 20, respectively. For each subepoch, 1,000 segments were extracted from up to 50 cases.

The learning rate was decreased exponentially and the momentum linearly increased. The values that had to be reached at the last epoch were 10^{-4} for the former and 0.9 for the latter. The learning rate, initially set to 10^{-3}, started to lower at epoch 1. Updating learning rates through training is a way of making sure that convergence is reached and in a reasonable time (Jacobs, 1988; Zeiler, 2012). The learning optimizer was RmsProp (Tieleman and Hinton, 2012), with $\rho = 0.9$ (decay rate) and $\epsilon = 10^{-4}$ (smoothing term that avoids divisions by zero). RmsProp was combined with Nesterov momentum (Nesterov, 1983), as proposed by Sutskever et al. (2013). The momentum value was set to $m = 0.6$ and normalized. Additionally, weight decay was also implemented, in the form of L1 and L2 normalization with values $L1 = 10^{-6}$ and $L2 = 10^{-4}$, respectively.

Also, two "online" (done during training) data augmentation techniques were set by default. The first simply involved reflecting images with a 50% probability with respect to the X axis (from left to right). The second consisted in altering the mean and standard deviation of the images, following the next equation:

$$I' = (I + s) * m, \tag{1}$$

where s (shift) and m (multi) are drawn from Gaussian distributions of $(\mu = 0, \sigma = 0.05)$ and $(\mu = 1, \sigma = 0.01)$, respectively.

Finally, due to memory limitations, only three out of the six available channels were used to train the model, namely ADC, MTT, and rCBF. In some experiments, rCBF was replaced by rCBV. Only two segmentation classes were considered, foreground, representing the lesion, and background, representing everything else.

2.3. Experiments

To evaluate the use of enhancing learning techniques for identifying ischaemic stroke lesions in perfusion imaging data, six experiments were run (i.e., E0–E5) by varying one aspect of the model at a time, such as the type of data or other parameters. This was done in the form of a pipeline, performing pair-wise comparisons. At each stage of the pipeline, two models, with and without a particular change, were compared. The best performing model of each pair-wise comparison proceeded to the next stage, until the best performing model of all experiments was found.

To assess the performance of an experiment, k-fold cross-validation was employed, where $k = 5$. Cross-validation is essential to give a good estimate of the real performance of an experiment. If cross-validation hadn't been used, results would have highly depended on the composition of easy/hard cases in each set. For example, if the test set had only been made of easy cases, the performance achieved would have been greater that if they had been difficult cases. Overall, this not only increases the robustness of the results but also the confidence of the decisions related to the changes that have worked best.

2.3.1. Data Pre-processing

Performing adequate pre-processing of the data is essential to maximize the performance of the model. Some of the necessary pre-processing steps were already done by the ISLES organizers, such as co-registering all images per subject setting them to have the same dimension, also per subject, and removing extracranial tissues.

Additional pre-processing involved resampling all images to isotropic (i.e., 1 x 1 x 1 mm) voxels size, generating intracranial volume masks and normalizing the data to have zero mean and unit variance. The latter is strongly suggested by DeeMedic's creator as it would substantially affect performance. The intracranial volume masks were generating binarizing the TTP images, and applying binary dilation before the resampling to improve the boundaries. Due to memory constraints, all images had to be downsampled with a factor of 0.7 so they could fit in memory (Algorithm 1).

Algorithm 1 Data Pre-processing

```
Initialize dF = 0.7
for each subject do
    for each channel do
        resampled_channels ← resample(channel)
    end for
    mask ← compute_mask(channels)
    mask ← resample(mask)
    save_image(mask)
    for each resampled_channel do
        img ← normalize(resampled_channel, mask)
        save_image(img)
    end for
end for
```

2.3.2. E0 - Baseline Configuration

This experiment (i.e., E0) consisted in training the DeepMedic configuration described previously, with the default parameters using the pre-processed data. It established the baseline results. All future experiments were compared against this or a better performing one. The imaging modalities used as input channels were ADC, MTT, and rCBF.

2.3.3. E1 - Data Augmentation

We applied the data augmentation method known as intensity variance. It consists in randomly altering the intensity values within the Region of Interest (ROI) or GT region following a Gaussian distribution of mean and variance equal to the ones computed from the intensity values within the region.

The rationale behind this idea was to try to deal with one of the many complications of detecting the ischemic stroke lesion in these types of images: their intensity inhomogeneity. As mentioned by Maier et al. (2017), the intensity values within the lesion territory can vary significantly. By using a mean and variance based on the already available data, the intensities, while being different from the original, should not be too different so as the lesion is no longer recognizable.

This augmentation was done offline, which means that the altered subjects were created and saved to be fed to the network during training. It was decided to do it this way so as to avoid modifying DeepMedic's core code, which would in turn become very time consuming. Each new subject is a "clone" of the original, except for the intensity values within the ROI or GT label. All channels had their intensity modified. Algorithm 2 shows how this was done.

Algorithm 2 Data augmentation

```
Initialize clones_number = 1
for each subject do
    Load label
    for each clones_number do
        Initialize clone_path
        for each channel do
            roi ← channel[nonzero(label)]
            channel[nonzero(label)]
            ← gaussian(mean(roi), std(roi))
            save_image(channel, clone_path)
        end for
    end for
end for
```

This experiment used the same baseline configuration parameters as E0, with the exception that the data had been augmented. The original 43 subjects had been "cloned," following the procedure described above. Thus, the total number of available training subjects became 86. However, since validation or testing in augmented subjects is meaningless, only the subjects inside the training set contained clones. Naturally, clones of the validation and test subjects were not part of the training set.

2.3.4. E2 - Transfer Learning With Error Maps

The goal of this experiment was to improve the performance of a pre-trained model (i.e., the best performing model so far), by fine-tuning the model with its error maps (i.e., weighted maps), using them to draw more image segments from difficult regions (i.e., those where errors were bigger).

Fine-tuning is a type of transfer-learning aimed at improving the performance of a network pre-trained for a different - although similar- task to the one the model was originally trained for (Pan et al., 2010). For example, two different tasks can have the same goal and only vary on the information that is provided to complete them. Usually, this technique involves re-training a network while "freezing" the first layers, meaning that their parameters (weights) are kept fixed during training. Each consecutive layer of a CNN generates more complex features from the ones detected in the previous layer. Consequently, the first layers contain simpler features that are common for similar problems, and thus can be "transferred" to a similar task. Then, new data is used to retrain the final

layers, tuning the network to improve performance on the new task.

In other words, the aim of fine-tuning is to adapt the network to the small details that make the new task different, which means the learning rate has to account for that by being considerably small compared to the original rate the model was pre-trained with. For that reason, while the learning rate of the initial model was initialized to 10^{-3}, the rate for this experiment was $5x10^{-4}$. There are three possible benefits of using transfer learning: a higher start, a higher slope and a higher asymptote (Aytar and Zisserman, 2011). When performing transfer learning, it's possible that one, two, all or none of these benefits appear.

To improve learning, an adaptive sampling method has been proposed (Berger et al., 2017) for DeepMedic. It consists in extracting more image patches in the regions where the prediction error is bigger, according to error maps generated throughout training. DeepMedic already offers the possibility of using weighted maps for the sampling process, which essentially serves the same function but in a static way (i.e., maps must be generated beforehand and are not updated during training). By using these maps, image segments are extracted more often from those regions where the weights are bigger. Error maps, one per subject and class, were obtained by computing the square error between each voxel of the GT label and the predicted probability map. The probability maps were obtained from the segmented test cases of each fold, meaning that the error maps for all subjects could be computed. These maps were normalized to zero mean and unit variance for homogeneity between subjects.

The paths of the computed error maps were included in different files, one for each class. These files were specified in the configuration parameters, each line representing a subject, which had to be coherent between files. Weighted maps can be defined both for training and validation. Since the goal was to improve the network performance, only error maps for the training cases were provided. In these cases, fine-tuning was performed by retraining the best model so far while extracting more image segments in those regions where errors where bigger, with the aid of pre-computed error maps. All convolutional layers were left frozen, thus only tuning the FC layers.

2.3.5. E3, E4, and E5 - Transfer Learning With rCBV

Perfusion parametric maps rCBF and rCBV display different appearance depending on the area under consideration. In the core of the stroke both sequences have substantially low values. However, in the penumbra (i.e., affected but salvageable region), while rCBF is slightly reduced, rCBV can be normal or even have higher values compared to normal tissue. Both sequences have been used to segment the stroke (Chen and Ni, 2012).

In this experiment, the best performing model so far is retrained using the ADC, MTT, and rCBV as input channels. Recall that until now, models have used the ADC, MTT, and rCBF as input channels for training, as defined in the baseline configuration.

The goal of E3 is to make predictions more robust by tuning the weights of the FC layers, similar to experiment E2 in previous section. This would make the network more sensitive to small changes between rCBF and rCBV, which can be crucial to accurately segmenting the stroke.

E4 and E5 are essentially the same as E3 with the exception of the number of frozen layers. E4 has only the first four convolutional layers frozen, whereas E5 has no frozen layers at all. This is useful to also examine the effect of freezing different numbers of layers for the lesion segmentation task.

2.4. Post-processing

In order to test whether the predictions of DeepMedic could be further improved, different post-processing techniques were implemented, based on threshold tuning the DeepMedic's probability output and performing binary morphological operations in the binarized result.

However, before applying any of these techniques, DeepMedic outputs (i.e., predicted lesion and class probability maps) had to be resampled to their corresponding subjects' original image space so that results could be interpreted in the same dimensional space as the original data. Hence, we resampled all outputs per subject using the inverse affine transformation applied to the original images in the ISLES 2017 dataset.

2.4.1. Threshold Tuning

After computing the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves it is possible to obtain the optimal threshold to be applied to the DeepMedic probabilistic output, which maximizes the desired metrics. To this end, we implemented two threshold tuning procedures, one for each curve. It is worth noting that both methods were independent and their results were not combined. Also, both curves were computed using the *Scikit-learn* library.

The first threshold tuning procedure, Threshold Tuning 0 (THT0), consisted in obtaining the point where ($precision * recall$) was maximum. This is the furthest point from the bottom-left corner and thus returns the maximum value for the DSC metric. To compute it, we concatenated the original GT and the probability map of the foreground class of all subjects (separately) to compute the curve, and, then, selected the optimal threshold.

The second procedure, Threshold Tuning 1 (THT1), based on the ROC curve, consisted in obtaining the point where ($TruePositiveRate(TPR) - FalsePositiveRate(FPR)$) was maximum. This represents the furthest point from the bottom-right corner and thus the optimal threshold, giving the maximum value for the Bookmaker Informedness (BM) metric. Again, all subjects' labels and probability maps were concatenated to compute the curve, and, then, select this threshold.

The goal of both procedures was to obtain the best average threshold for the results from the validation set to apply it to the test set. This was done for all folds independently. This guarantees that the tuning is not performed on the test (i.e., validation) cases, which accounts for a real scenario where the GT for the test cases are not available.

2.4.2. Binary Morphological Operations

Binary morphological operations are mathematical operations used to modify shapes in binary images through a structuring element: a shape to probe the image. Closing is a binary morphological operation that can fill holes in big predicted lesions or join reasonably close small ones to make predictions more robust. It combines two other simpler morphological operations: dilation, which expands shapes in an image, and erosion, which shrinks them. In both cases, the center of the structuring element is placed at every pixel of the image and a decision is made. In the case of dilation, a pixel is set to 1 if there are any pixels equal to one within the shape of the structuring element, otherwise it's set to zero. Erosion performs the exact opposite operation, a pixel is set to 0 as long as there is any pixel of value 0 within the area covered by the structuring element.

Furthermore, there are two decisions to make regarding this operation: the shape and size of the structuring element and the number of iterations. While the first determines the final output and thus the goodness of the prediction, the second defines the number of times that the *dilation* operation inside the *close* function is repeated (followed by the same number of iterations for the *erosion* operation) [2].

After few experiments, the optimal structuring element was a 3D ball with a radius of 3 voxels, whereas the number of iterations was tuned by selecting the average of the ones that achieved the maximum DSC score on validation cases. This post-processing step was named Filling Holes (FH).

2.5. Evaluation

At each state of the post-processing pipeline, multiple performance metrics were computed to compare the predicted segmented lesions with the GT. These metrics were TPR, True Negative Rate (TNR), Positive Predictive Value (PPV), Accuracy (ACC), DSC, Matthews Correlation Coefficient (MCC), and Hausdorff Distance (HD). Being True Positives (TP) the voxels predicted to be positives and identified positives by the configuration evaluated, True Negatives (TN) the voxels predicted to be negatives and identified negatives, False Positives (FP) the voxels predicted to be negatives but identified positives and False negatives (FN),the voxels predicted to be positives but identified negatives, these metrics are defined as follows:

- **TPR:** Also known as *sensitivity* or *recall*, measures the rate of true positives with respect to the number of real positive cases.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2}$$

- **TNR:** Also known as *specificity*, measures the rate of true negatives with respect to the number of real negative cases.

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{3}$$

[2]https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.ndimage.morphology.binary_closing.html

- **PPV:** Also known as *precision*, measures the proportion of true positives with respect to all predicted positives.

$$PPV = \frac{TP}{P'} = \frac{TP}{TP + FP} \tag{4}$$

- **ACC:** Is a measure of statistical bias. Represents how close the predictions are from the true values.

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

- **DSC:** The Dice similarity coefficient measures the harmonic mean of PPV and TPR. (Landis and Koch, 1977) define the intervals and the associated "strength of agreement": [< 0.00] (Poor), [$0.00 - 0.20$] (Slight), [$0.21 - 0.40$] (Fair), [$0.41 - 0.60$] (Moderate), [$0.61 - 0.80$] (Substantial), [$0.81 - 1.00$] (Almost perfect).

$$F_i = 2 * \frac{PPV * TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN} \tag{6}$$

- **MCC:** Also known as the *phi coefficient* or Matthews correlation coefficient, is considered a balanced metric of the quality of binary classification, thus robust to class imbalance. Values range from –1 (perfect negative correlation) to 1 (perfect positive correlation), being 0 equal to random prediction. This metric is considered to be the most meaningful, specially for imbalanced data (Chicco, 2017).

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

- **HD:** Measures the distance between two subsets. A_S and B_S are equivalent to P (real true cases) and P′ (predicted true cases), and $d(\cdot)$ is the euclidean distance between two points.

$$HD(A_s, B_s) = \max\{\max_{a \in A_s} \min_{b \in B_s} d(a, b), \max_{b \in B_s} \min_{a \in A_s} d(b, a)\} \tag{8}$$

Since we used k-fold cross-validation, these metrics were averaged per fold and also between folds. This means that performance metrics were available per subject (both for the validation and test sets' subjects of every fold), per fold and per experiment. Performance curves, known as precision PPV vs. recall TPR, error bar and Bland-Altman (Bland et al., 1986) plots were also produced. In addition, the DeepMedic plotting script was slightly modified to generate the progress of metrics such as accuracy or DSC on training and validation sets through the different epochs.

3. RESULTS

3.1. Segmentation Performance During Training

The segmentation performance for validation and training sets during the training process is shown in **Figure 1**. The DSC coefficient was stable after improving during few epochs. On the other hand, sensitivity (i.e., TPR) improved at first but then

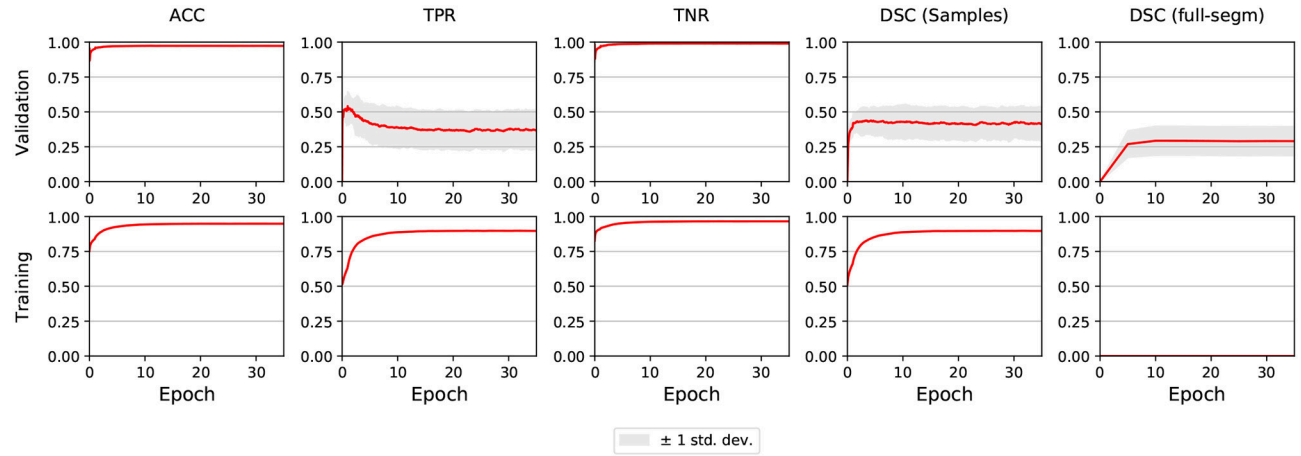

FIGURE 1 | E0 - Segmentation metrics of validation and train subjects during training. The graphs shown are the averages of all 5 folds. The light gray area illustrates ±1 standard deviation. Full segmentation on training cases was not performed by DeepMedic, reason why the lower-right graph is empty.

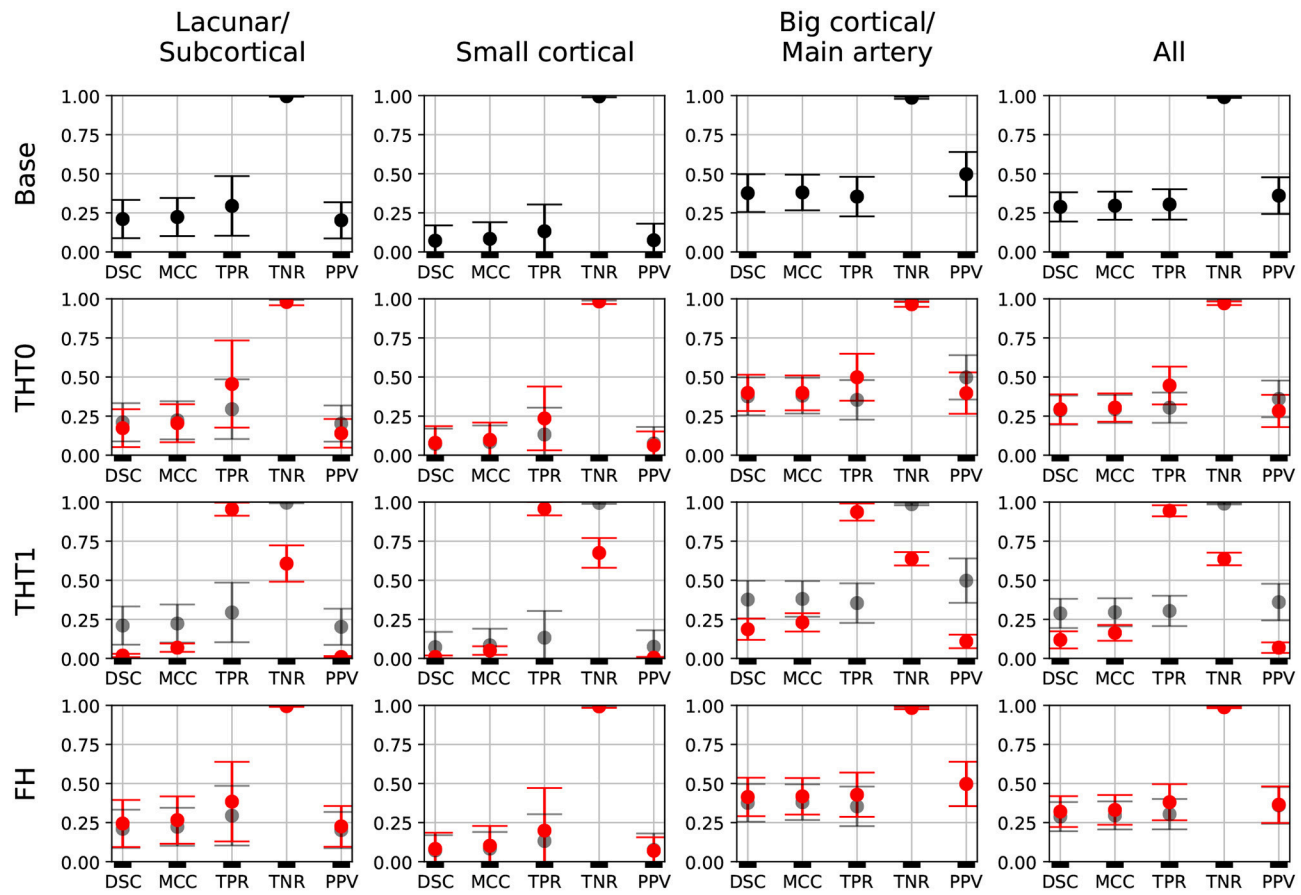

FIGURE 2 | E0 - Error bars. Each metric for each post-processing step and lesion category is presented. A fourth column, representing all subjects, is also included. Base results are present in all rows for comparison purposes (semi-transparent black). Each marker represents the mean value, and the upper and lower limits represent the 95% confidence interval. The metrics shown are: Dice similarity coefficient (DSC), Matthews correlation coefficient (MCC), True positive rate (TPR), True negative rate (TNR), and Positive predicted value (PPV).

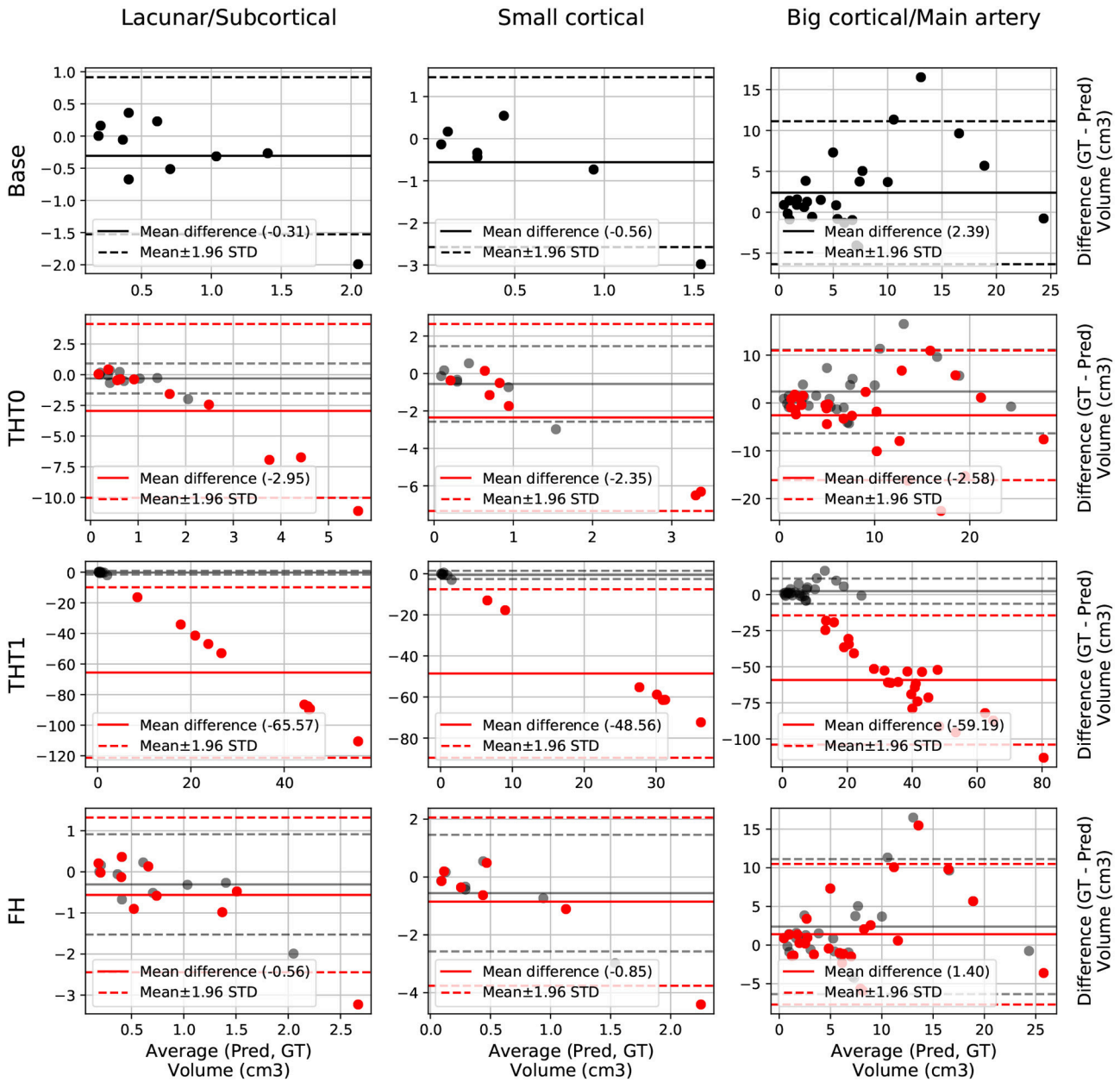

FIGURE 3 | E0 - Volume Bland-Altman analysis. Each lesion category (lacunar/subcortical, small cortical and big cortical) and post-processing step (THT0, THT1, FH, and base) are included. Base is also included in each post-processing step row for comparison purposes (semi-transparent black). Each point represents one subject. The solid line is the mean difference, whereas the dotted-line represents the limits of agreement, computed as mean±1.96 STD. The x axis is the average volume between the predicted segmentation and the ground truth, whereas the y label is the difference.

worsened and remained stable. Mean accuracy and specificity, while being very high, did not account for the imbalanced nature of the data.

In E1, sensitivity took more time to reach its peak compared to E0, but when it stabilized the asymptote was slightly higher. Also, while DSC behaved similarly to E0, it also achieved higher values. In E2-E5, the metrics for the first epoch had the same value as for the last epoch in E1, and did not improve throughout the training process.

3.2. Baseline Segmentation Performance

Figure 2, shows the error bars for each metric, post-processing step and lesion category for E0. TPR was highly variable for small stroke lesions, regardless of whether they were lacunar or cortical, especially after the THT0 and FH post-processing steps. THT1 produced consistently worse results in terms of accuracy for small stroke lesions, despite achieving higher TPR (i.e., sensitivity). The segmentation of big cortical/main artery stroke lesions was considerably better than those for the other stroke subtypes.

The Bland-Altman plot showing the volumetric agreement between the GT and the results from E0 after each post-processing step can be seen in **Figure 3**. THT1 produced the worst results in terms of volumetric agreement regardless of the stroke subtype, considerably inflating the stroke lesion volume. This method for selecting the optimal threshold for binarizing the probabilistic stroke lesion maps obtained, overestimated the stroke lesion size in general. This overestimation, reflected in the difference between the volumes of the GT and the output from applying THT1, was proportional to the stroke lesion size. The post-processing step pf FH slightly improved the volumetric estimation of big cortical strokes with respect to the base measurements.

3.3. Experiments' Results

E1 was the best performing model, with an average DSC of 0.34 after applying FH. This proves the efficacy of using the data augmentation method selected (i.e., intensity variance). It also proves the importance of performing post-processing tasks, such as THT0 and FH, instead of simply focusing on pre-processing and then relying on the output of the network.

Table 1 and **Figure 4** contain a summary of all experiments. E1 was superior to E0 and the rest experiments yielded results close to E1, but they were not able to improve it. E4 and E5 are not shown because their results were very similar to E3 but slightly inferior. In general, the transfer learning approaches (E2-E5) evaluated did not improve the accuracy in the results.

Table 1 shows the key metrics of each experiment both for all post-processing steps. On average, FH performed best. PPV and consequently DSC were the metrics that determined the best performing model.

Figure 4 depicts the DSC error bars for all post-processing steps and lesion categories. Big cortical lesions were easier to segment than the rest (i.e., small lesions).

Additionally, **Figure 5** shows the precision-recall curves for all experiments. Results are very different depending on the cases that fall in each fold. This is a clear sign of the heterogeneous nature of the data and the inability of the network to generalizing well. Also from these graphs, results from E1 are slightly superior to E0 and similar to E2. Interestingly, while E3 produced the worst results, its predictions were the least heterogeneous (i.e. the curves are more closer to each other than in any other experiment).

The winner (Choi et al., 2017) of the ISLES 2017 challenge, achieved 0.31 DSC and 103.64 HD when the final results were published in September of 2017, but since then the challenge has remained open. Consequently, more participants have joined the challenge and the current top performer, as of the time of writing this manuscript, achieved 0.36 DSC and 29.37 HD.

To perform a fair comparison between our E1 and the current state of the art performance, E1 was retrained using all train data for training and tested on the unlabeled test set of the challenge. FH was then applied to the predicted lesions using the average number of iterations in E1 and the results uploaded to the SMIR web page[3].

[3]www.smir.ch

TABLE 1 | Summary of the main metrics for all experiments (i.e., E0-E3).

	Post-proc	DSC	HD	MCC	TPR	TNR	PPV
E0	Base	0.29	62.22	0.30	0.30	0.99	0.36
	THT0	0.29	72.83	0.30	0.45	0.97	0.28
	THT1	0.12	99.62	0.16	0.94	0.64	0.07
	FH	0.32	59.47	0.33	0.38	0.99	0.36
E1	Base	0.32	49.89	0.32	0.34	0.99	0.38
	THT0	0.31	72.33	0.33	0.49	0.97	0.30
	THT1	0.13	100.29	0.18	0.96	0.65	0.07
	FH	0.34	47.85	0.35	0.40	0.99	0.39
E2	Base	0.31	48.48	0.32	0.34	0.99	0.38
	THT0	0.31	71.42	0.33	0.48	0.97	0.30
	THT1	0.13	100.19	0.18	0.96	0.68	0.08
	FH	0.33	46.74	0.35	0.40	0.99	0.38
E3	Base	0.30	57.37	0.31	0.36	0.99	0.36
	THT0	0.31	66.37	0.32	0.42	0.98	0.32
	THT1	0.12	99.94	0.17	0.97	0.63	0.07
	FH	0.33	53.94	0.34	0.42	0.99	0.36

Average metrics from the base prediction and all post-processing steps are shown. These are: Threshold tuning 0 (THT0), Threshold tuning 1 (THT1) and Filling holes (FH). The metrics shown are: Dice similarity coefficient (DSC), Hausdorff distance (HD), Matthews correlation coefficient (MCC), True positive rate (TPR), True negative rate (TNR), and Positive predicted value (PPV).

E1 achieved 0.29 DSC and 49.75 HD on the test set, as reported by the SMIR web page. This value is inferior to the 0.34 DSC achieved in the E1 experiment and also to the current first position of the challenge. This difference could be because of the fact that either the network or the number of iterations for FH computed in E1 were not able to generalize well on the test data.

3.4. Visual Evaluation of the Results

Figures 6–8 show the results from E1 for representative axial slices superimposed in the ADC image, from three subjects randomly selected from each category. In general, stroke lesion predictions were better in E1, but not by a large margin, and these figures, overall, exemplify the results obtained.

Compared to E0, some cases were better segmented, but this was not always the case. For example, the stroke lesion prediction for subject 9 (lacunar infarct) achieved a DSC score of 0.45 in E0, whereas in E1 it achieved 0.56. However, for subject 21 (small cortical infarct), the DSC score for E0 was 0.26, whereas in E1 it was 0.24, i.e., a slightly worse score. In general, E1's DSC was 10.34% better than E0's and 6.25% for FH. Most results were visually very similar. Also, in E1, post-processing steps (i.e., THT0, THT1, FH) did not improve results as much as they did in E0.

The GT, obtained from the structural T2-weighted images, not always includes the whole regions with restricted diffusion (i.e., dark regions in the ADC map). Contrastingly, in cases of large strokes, it includes the cerebrospinal fluid in the sulci. For cases in which the GT extent agrees with the region of restricted diffusion, the results are better (e.g., cases 9 and 32).

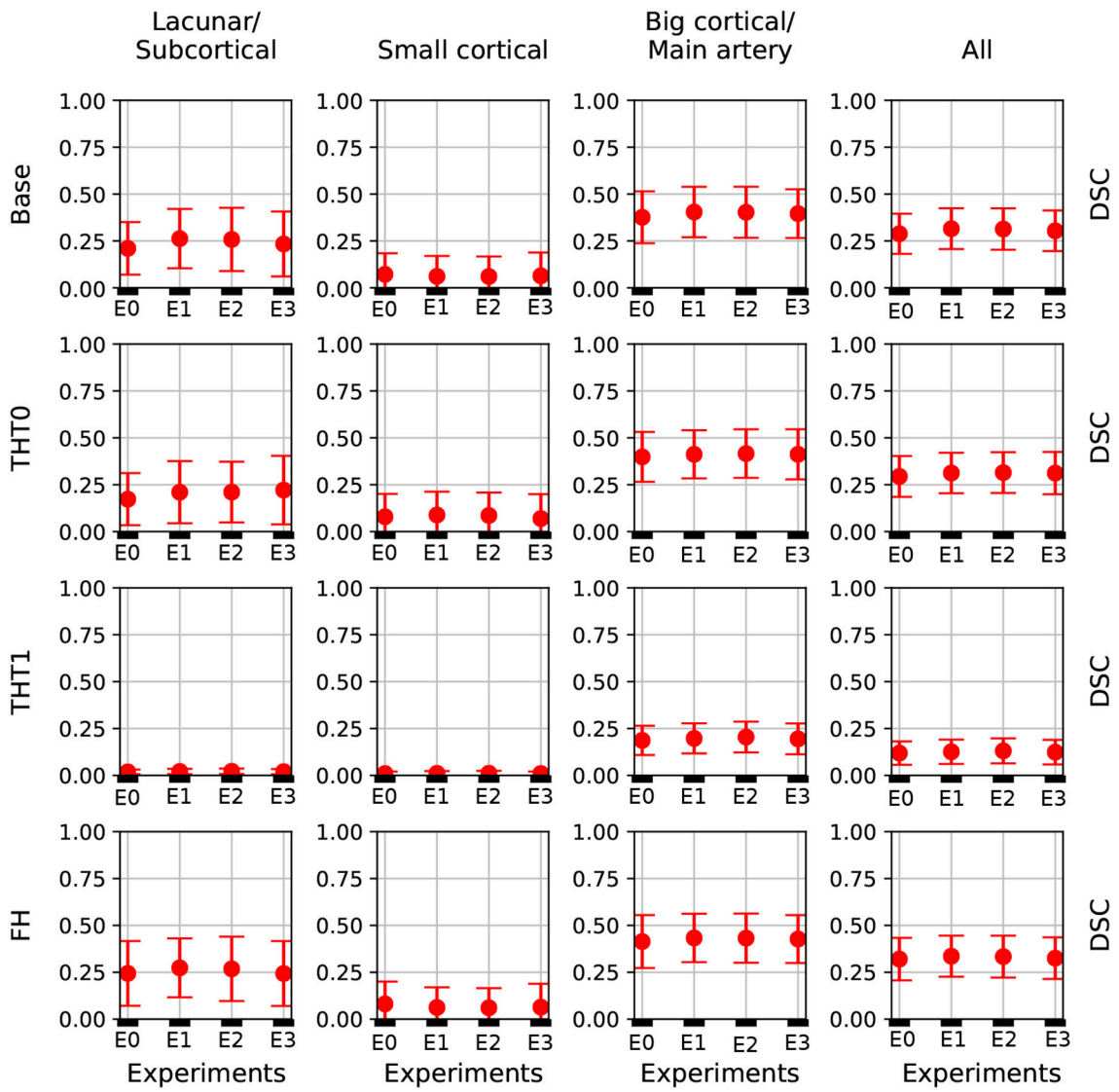

FIGURE 4 | DSC error bars of all experiments for the base prediction and FH and each lesion category.

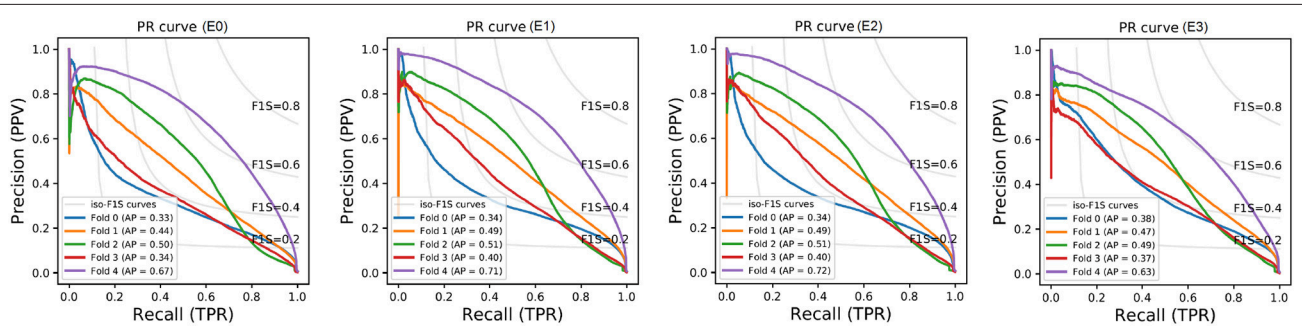

FIGURE 5 | Performance curves of E0-E3. The gray lines indicate the iso-F1S curves, the value of DSC for each point in the graph. The AP metrics are also included.

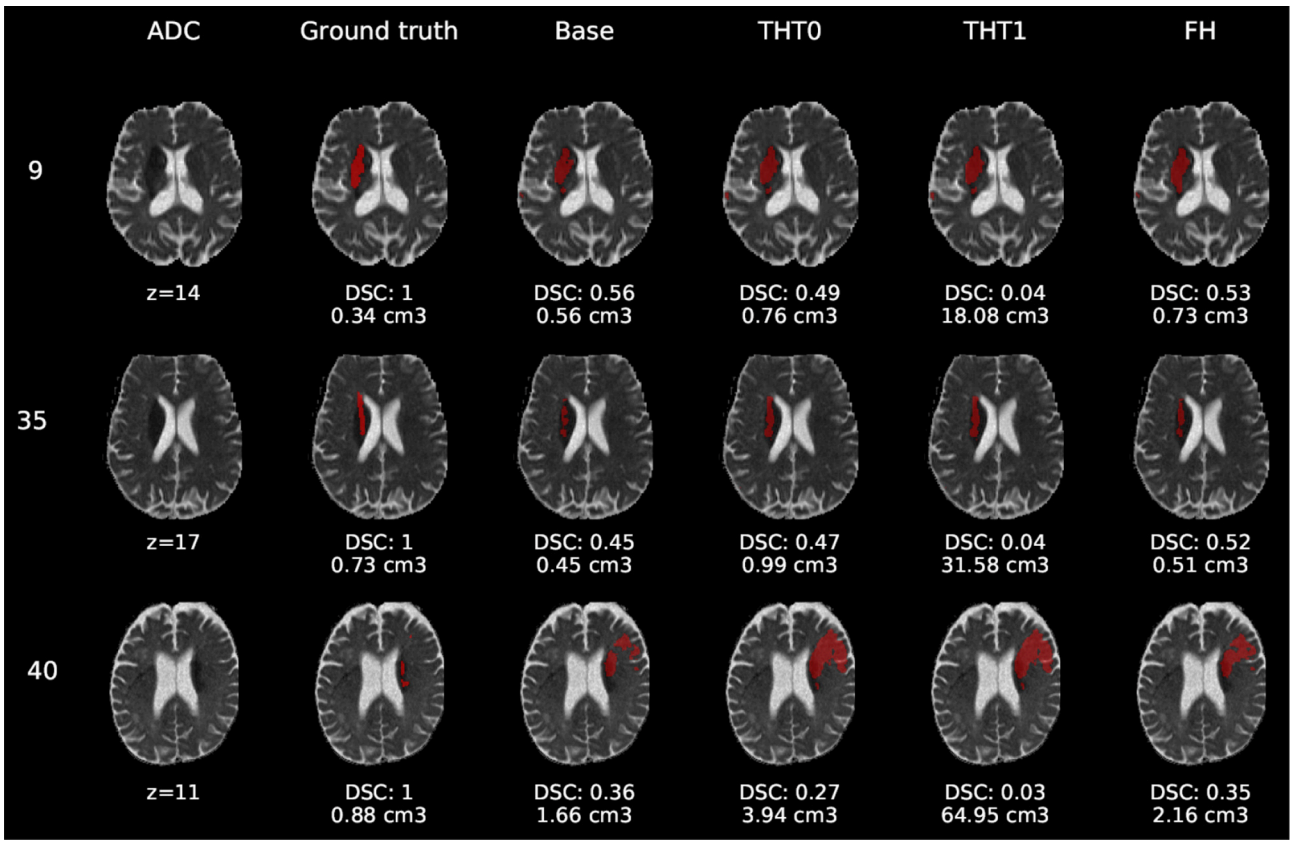

FIGURE 6 | E1 - Visual segmentation comparison of lacunar/subcortical lesions. The examples include the predicted lesions after each post-processing step. Images are 2D slices, their cut coordinate in the z axis is included, as well as the volume of each segmentation and the DSC achieved.

Visually, results obtained applying THT1 to the DeepMedic's output does not appear to be disparately wrong compared to those obtained applying THT0 and/or FH.

4. DISCUSSION

Although the best enhanced learning strategy proposed (FH) improved the segmentation results in the majority of cases, our results are still suboptimal. We used the default configuration, batch size, learning rate and activation functions of a CNN scheme designed to segment tumors from structural MRI sequences. Also, instead of pre-training the network with data of similar nature, but a varied, larger dataset, and fined-tune it with this ISLES 2017 dataset, we directly trained it with a subset from the latter. Therefore, overfitting was still a problem even with data augmentation. Reducing it could be achieved by modifying the number of layers and the size of kernels, and thus the number of network parameters. It could also be remedied by using data from other challenges, or even past iterations of ISLES that also contain the same sequences for segmenting the stroke lesion. Moreover, the learning rate schedule should lower the learning rate at predefined epochs. We used the DeepMedic's default without prior training the model to determine when it would be more convenient to lower the learning rate, and the schedule was set to exponential decrease. Future work should try to lower the learning rate only when necessary.

In addition to these suggestions, data augmentation for medical images can also be done by employing Generative Adversarial Networks (GAN), which have recently being used in multiple works (Yi et al., 2018). For example, (Shin et al., 2018) employs GANs not only to improve accuracy of deep learning segmentation models through the generation of synthetic brain tumors, but also to achieve subject anonymity. Other works have also applied this technique for detecting brain metastases (Han et al., 2019), classifying liver lesions (Frid-Adar et al., 2018) and for other medical segmentation tasks (Bowles et al., 2018). Overall, these works agree on the performance improvements achieved by applying data augmentation based on GANs.

Despite the limitations previously mentioned, the GT used should be put into question. As the examples selected show, it did not accurately cover the region of restricted diffusion in the ADC images, underestimating it mainly for small infarcts and overestimating large infarcts, including regions of fluid in the sulci. The GT was generated using the structural

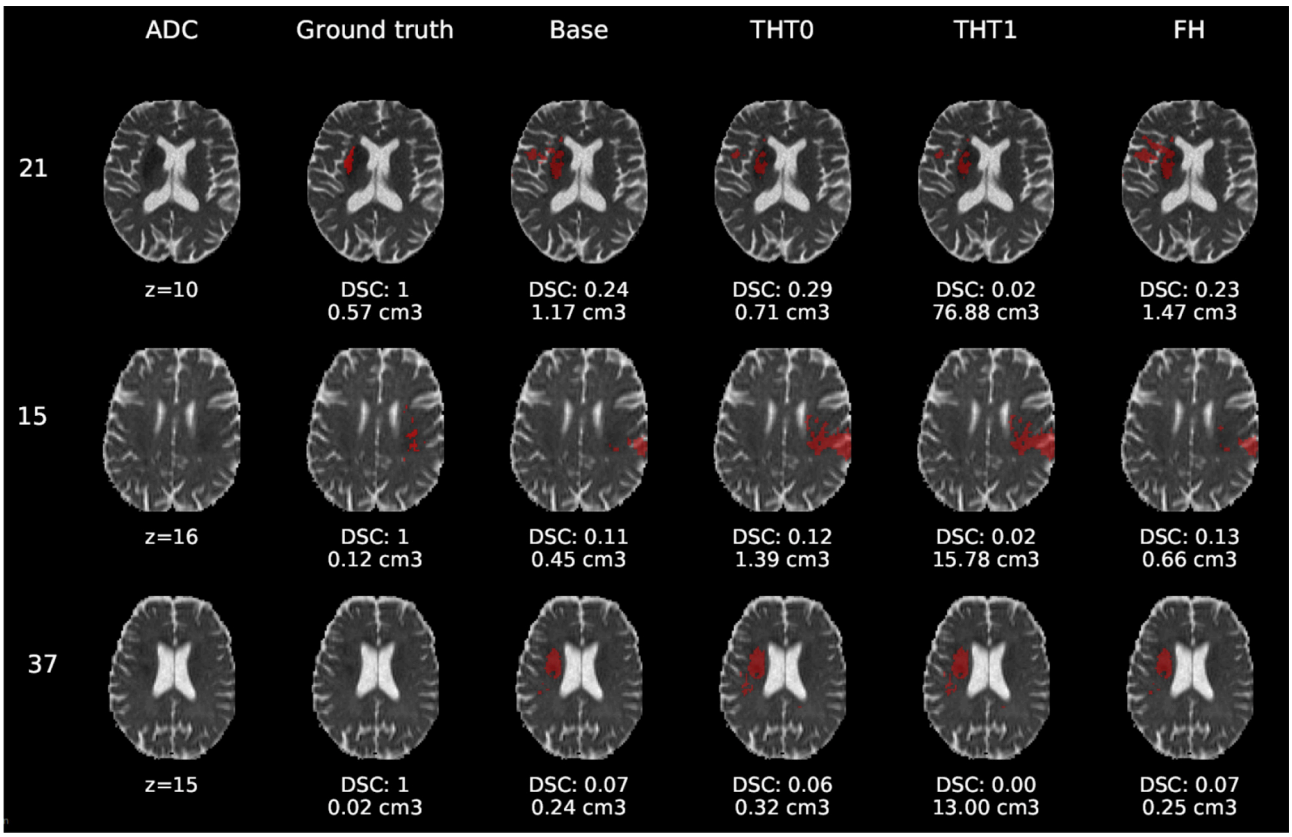

FIGURE 7 | E1 - Visual segmentation comparison of small cortical lesions. The examples include the predicted lesions after each post-processing step. Images are 2D slices, their cut coordinate in the z axis is included, as well as the volume of each segmentation and the DSC achieved.

T2-weighted images (i.e., including FLAIR), not provided. The mismatch between structural, diffusion and perfusion MRI modalities is well-known (Chen and Ni, 2012; Motta et al., 2015; Straka et al., 2010).

Precisely, the perfusion/diffusion mismatch has been reported to provide a practical and approximate measure of the tissue at risk, being used to identify acute stroke patients that could benefit from reperfusion therapies. Clinical studies also show that early abnormality on diffusion-weighted imaging can overestimate the infarct core by including part of the tissue "at risk," and the abnormality on perfusion weighted imaging overestimates this "at risk" tissue by including regions of benign tissue with reduced blood perfusion (Chen and Ni, 2012).

The diffusion/fluid attenuated inversion recovery (DWI/FLAIR) mismatch is also well-known. Together with the perfusion/diffusion mismatch it is recognized as an MRI marker of evolving brain ischaemia. A clinical trial that examined whether the DWI/FLAIR mismatch was independently associated with the diffusion/perfusion mismatch or not, concluded that in the presence of the latter, the DWI/FLAIR pattern could indicate a shorter time between the scan and the last time the tissue seen was normal (Wouters et al., 2015). The CNN scheme evaluated does not take into account the time from the stroke onset—information not provided.

Finally, the types of infarcts were not evenly represented in the dataset. The large cortical strokes were predominant, which could explain the bias in the results favoring the cases when the stroke was of this subtype. The involvement of personnel with relevant clinical knowledge in the generation of datasets to be used for developing algorithms aimed to clinical research would be advisable in the future.

5. CONCLUSION

The model that used data augmentation had the best performance, achieving an average DSC score of 0.34 for the test cases after applying FH. This was a reasonable outcome considering that the network clearly suffered from overfitting, for which data augmentation is a well-known remedy.

Also, of all post-processing steps evaluated, FH produced the best improvements on average over the base prediction by the network. The second best was THT0, which in some cases surpassed FH. The results from applying THT1, although worst in terms of accuracy, were not visually very different.

In summary, considering all the complications related to the nature of this problem that have been mentioned along this document, it is clear that much work is left to be done in order

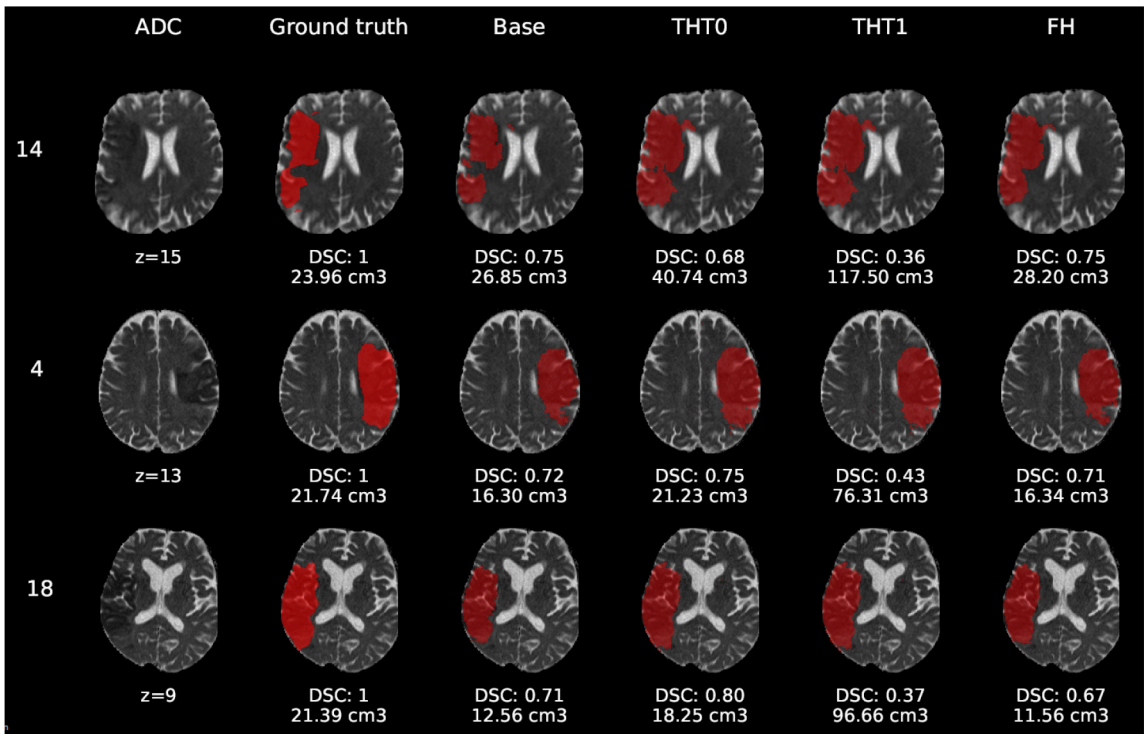

FIGURE 8 | E1 - Visual segmentation comparison of big cortical lesions. The examples include the predicted lesions after each post-processing step. Images are 2D slices, their cut coordinate in the z axis is included, as well as the volume of each segmentation and the DSC achieved.

to achieve reasonable results on the task of ischaemic stroke segmentation so that an automatic system can operate reliably on a clinical environment.

AUTHOR CONTRIBUTIONS

CPM, MCVH, MFR, and TK conceived and presented the idea. CPM and MCVH planned the experiments. CPM carried out the experiments. All authors provided critical feedback and analysis, and contributed to the manuscript.

FUNDING

This study was partially funded by the Indonesia Endowment Fund for Education (LPDP) of Ministry of Finance, Republic of Indonesia (MFR), Row Fogo Charitable Trust (Grant No. BRO-D.FID3668413)(MCVH), the European Union Horizon 2020 [PHC-03-15, project No 666881, SVDs@Target] (MCVH), and the UK Biotechnology and Biological Sciences Research Council (BBSRC) through the International Partnership Award BB/P025315/1 (MCVH).

REFERENCES

Aytar, Y., and Zisserman, A. (2011). "Tabula rasa: model transfer for object category detection," in *2011 IEEE International Conference on Computer Vision (ICCV)* (Barcelona: IEEE), 2252–2259. doi: 10.1109/ICCV.2011.6126504

Berger, L., Hyde, E., Cardoso, J., and Ourselin, S. (2017). An adaptive sampling scheme to efficiently train fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1709.02764*. doi: 10.1007/978-3-319-95921-4-26

Bland, J. M., and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1, 307–310.

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., et al. (2018). Gan augmentation: augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*.

Brosch, T., Tang, L. Y., Yoo, Y., Li, D. K., Traboulsee, A., and Tam, R. (2016). Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation. *IEEE Trans. Med. Imaging* 35, 1229–1239. doi: 10.1109/TMI.2016.2528821

Chen, F., and Ni, Y.-C. (2012). Magnetic resonance diffusion-perfusion mismatch in acute ischemic stroke: an update. *World J. Radiol.* 4:63. doi: 10.4329/wjr.v4.i3.63

Chen, L., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *CoRR. arXiv:1706.05587v3*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2018a). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 834–848. doi: 10.1109/TPAMI.2017.2699184

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. *arXiv preprint arXiv:1802.02611*. doi: 10.1007/978-3-030-01234-2-49

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Mining* 10:35. doi: 10.1186/s13040-017-0155-3

Choi, Y., Kwon, Y., Paik, M. C., and Joon, B. (2017). *Ischemic Stroke Lesion Segmentation With Convolutional Neural Networks for Small Data.* ISLES 2017 Challenge. Available online at: http://www.isles-challenge.org/ISLES2017/articles/choi.pdf

de Brebisson, A., and Montana, G. (2015). "Deep neural networks for anatomical brain segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 20–28. doi: 10.1109/CVPRW.2015.7301312

Fantini, S., Sassaroli, A., Tgavalekos, K. T., and Kornbluth, J. (2016). Cerebral blood flow and autoregulation: current measurement techniques and prospects for noninvasive optical methods. *Neurophotonics* 3:031411. doi: 10.1117/1.NPh.3.3.031411

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). "Synthetic data augmentation using gan for improved liver lesion classification," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (Boston, MA: IEEE), 289–293.

Ghafoorian, M., Karssemeijer, N., van Uden, I. W., de Leeuw, F.-E., Heskes, T., Marchiori, E., et al. (2016). Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. *Med. Phys.* 43 (Alexandria, VA: American Association of Physicists in Medicine), 6246–6258. doi: 10.1118/1.4966029

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., et al. (2017). "Transfer learning for domain adaptation in mri: application in brain lesion segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 516–524.

Glorot, X., and Bengio, Y. (2010). "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., et al. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage Clin.* 17, 918–934. doi: 10.1016/j.nicl.2017.12.022

Han, C., Murao, K., Satoh, S., and Nakayama, H. (2019). Learning More with less: GAN-based medical image augmentation. *arXiv e-prints* arXiv:1904.00838.

He, K., Zhang, X., Ren, S., and Sun, J. (2014). "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision* (Springer), 346–361. doi: 10.1007/978-3-319-10578-9-23

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago), 1026–1034. doi: 10.1109/ICCV.2015.123

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Netw.* 1, 295–307. doi: 10.1016/0893-6080(88)90003-2

Kamnitsas, K., Ledig, C., Newcombe, V. F., Simpson, J. P., Kane, A. D., Menon, D. K., et al. (2017). Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Med. Image Anal.* 36, 61–78. doi: 10.1016/j.media.2016.10.004

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). 'Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 1097–1105.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

López-Zorrilla, A., de Velasco-Vázquez, M., Serradilla-Casado, O., Roa-Barco, L., Graña, M., Chyzhyk, D., et al. (2017). "Brain white matter lesion segmentation with 2d/3d cnn," in *International Work-Conference on the Interplay Between Natural and Artificial Computation* (Springer), 394–403.

Lucas, C., and Heinrich, M. P. (2017). *2d Multi-Scale Res-Net for Stroke Segmentation*. ISLES 2017 Challenge. Available online at: http://www.isles-challenge.org/ISLES2017/articles/lucas.pdf.

Maier, O., Menze, B. H., von der Gablentz, J., Häni, L., Heinrich, M. P., Liebrand, M., et al. (2017). Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Med. Image Anal.* 35, 250–269. doi: 10.1016/j.media.2016.07.009

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)* (Stanford, CA: IEEE), 565–571.

Motta, M., Ramadan, A., Hillis, A. E., Gottesman, R. F., and Leigh, R. (2015). Diffusion–perfusion mismatch: an opportunity for improvement in cortical function. *Front. Neurol.* 5:280. doi: 10.3389/fneur.2014.00280

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate o (1/kˆ 2). *Dokl. Akad. Nauk SSSR* 269, 543–547.

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowledge Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Petrella, J. R., and Provenzale, J. M. (2000). Mr perfusion imaging of the brain: techniques and applications. *Ame. J. Roentgenol.* 175, 207–219. doi: 10.2214/ajr.175.1.1750207

Rachmadi, M. F., del C. Valdés-Hernández, M., Agan, M. L. F., Perri, C. D., and Komura, T. (2018b). Segmentation of white matter hyperintensities using convolutional neural networks with global spatial information in routine clinical brain mri with none or mild vascular pathology. *Comput. Med. Imaging Graph.* 66, 28–43. doi: 10.1016/j.compmedimag.2018.02.002

Rachmadi, M. F., del C. Valdés-Hernández, M., and Komura, T. (2018a). "Transfer learning for task adaptation of brain lesion assessment and prediction of brain abnormalities progression/regression using irregularity age map in brain mri," in *PRedictive Intelligence in Medicine*, Rekik, I., Unal, G., Adeli, E., and Park, S. H., editors (Cham: Springer International Publishing), 85–93.

Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., et al. (2014). "A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 520–527.

Roy, P. K., Bhuiyan, A., Janke, A., Desmond, P. M., Wong, T. Y., Abhayaratna, W. P., et al. (2015). Automatic white matter lesion segmentation using contrast enhanced flair intensity and markov random field. *Comput. Med. Imaging Graph.* 45, 102–111. doi: 10.1016/j.compmedimag.2015.08.005

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2013). Overfeat: integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*.

Shin, H.-C., Tenenholtz, N. A., Rogers, J. K., Schwarz, C. G., Senjem, M. L., Gunter, J. L., et al. (2018). "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," In *International Workshop on Simulation and Synthesis in Medical Imaging* (Springer), 1–11.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Steenwijk, M. D., Pouwels, P. J., Daams, M., van Dalen, J. W., Caan, M. W., Richard, E., et al. (2013). Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (knn-ttps). *NeuroImage Clin.* 3, 462–469. doi: 10.1016/j.nicl.2013.10.003

Straka, M., Albers, G. W., and Bammer, R. (2010). Real-time diffusion-perfusion mismatch analysis in acute stroke. *J. Magn. Resonan. Imaging* 32, 1024–1037. doi: 10.1002/jmri.22338

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. (2017). "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Springer), 240–248.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning*, 1139–1147.

Tieleman, T., and Hinton, G. (2012). Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. *COURSERA Neural Netw. Mach. Learn.* 4, 26–31.

Van Nguyen, H., Zhou, K., and Vemulapalli, R. (2015). "Cross-domain synthesis of medical images using efficient location-sensitive deep network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer), 677–684.

Van Opbroek, A., Ikram, M. A., Vernooij, M. W., and De Bruijne, M. (2015). Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans. Med. Imaging* 34, 1018–1030. doi: 10.1109/TMI.2014.2366792

Wouters, A., Dupont, P., Ringelstein, E. B., Norrving, B., Chamorro, A., Grond, M., et al. (2015). Association between the perfusion/diffusion and diffusion/flair

mismatch: data from the axis2 trial. *J. Cereb. Blood Flow Metabol.* 35, 1681–1686. doi: 10.1038/jcbfm.2015.108

Xu, Y., Géraud, T., and Bloch, I. (2017). "From neonatal to adult brain mr image segmentation in a few seconds using 3d-like fully convolutional network and transfer learning," in *Image Processing (ICIP), 2017 IEEE International Conference on* (IEEE), 4417–4421. doi: 10.1109/ICIP.2017.8297117

Yi, X., Walia, E., and Babyn, P. (2018). Generative adversarial network in medical imaging: a review. *arXiv preprint arXiv:1809.07294*.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 27 June 2019
doi: 10.3389/fnagi.2019.00150

Dilated Saliency U-Net for White Matter Hyperintensities Segmentation Using Irregularity Age Map

Yunhee Jeong[1], *Muhammad Febrian Rachmadi*[1,2], *Maria del C. Valdés-Hernández*[2]* *and Taku Komura*[1]

[1] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom, [2] Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom

White matter hyperintensities (WMH) appear as regions of abnormally high signal intensity on T2-weighted magnetic resonance image (MRI) sequences. In particular, WMH have been noteworthy in age-related neuroscience for being a crucial biomarker for all types of dementia and brain aging processes. The automatic WMH segmentation is challenging because of their variable intensity range, size and shape. U-Net tackles this problem through the dense prediction and has shown competitive performances not only on WMH segmentation/detection but also on varied image segmentation tasks. However, its network architecture is high complex. In this study, we propose the use of Saliency U-Net and Irregularity map (IAM) to decrease the U-Net architectural complexity without performance loss. We trained Saliency U-Net using both: a T2-FLAIR MRI sequence and its correspondent IAM. Since IAM guides locating image intensity irregularities, in which WMH are possibly included, in the MRI slice, Saliency U-Net performs better than the original U-Net trained only using T2-FLAIR. The best performance was achieved with fewer parameters and shorter training time. Moreover, the application of dilated convolution enhanced Saliency U-Net by recognizing the shape of large WMH more accurately through multi-context learning. This network named Dilated Saliency U-Net improved Dice coefficient score to 0.5588 which was the best score among our experimental models, and recorded a relatively good sensitivity of 0.4747 with the shortest training time and the least number of parameters. In conclusion, based on our experimental results, incorporating IAM through Dilated Saliency U-Net resulted an appropriate approach for WMH segmentation.

Keywords: white matter hyperintensities, irregularity age map, saliency U-Net, MRI, segmentation, dilated convolution, deep learning

OPEN ACCESS

Edited by:
Andres Ortiz,
University of Málaga, Spain

Reviewed by:
Ruxandra Stoean,
University of Craiova, Romania
Jorge Munilla,
University of Málaga, Spain

***Correspondence:**
Maria del C. Valdés-Hernández
m.valdes-hernan@ed.ac.uk

Received: *30 January 2019*
Accepted: *07 June 2019*
Published: *27 June 2019*

Citation:
Jeong Y, Rachmadi MF, Valdés-Hernández MdC and Komura T (2019) Dilated Saliency U-Net for White Matter Hyperintensities Segmentation Using Irregularity Age Map. Front. Aging Neurosci. 11:150. doi: 10.3389/fnagi.2019.00150

1. INTRODUCTION

White matter hyperintensities (WMH) are commonly identified as signal abnormalities with intensities higher than other normal regions on the T2-FLAIR magnetic resonance imaging (MRI) sequence. WMH have clinical importance in the study and monitoring of Alzheimer's disease (AD) and dementia progression (Gootjes et al., 2004). Higher volume of WMH has been found in brains

of AD patients compared to age-matched controls, and the degree of WMH has been reported more severe for senile onset AD patients than presenile onset AD patients (Scheltens et al., 1992). Furthermore, WMH volume generally increases with the advance of age (Jagust et al., 2008; Raz et al., 2012). Due to their clinical importance, various machine learning approaches have been implemented for the automatic WMH segmentation (Admiraal-Behloul et al., 2005; Bowles et al., 2017).

Limited One-Time Sampling Irregularity Map (LOTS-IM) is an unsupervised algorithm for detecting tissue irregularities, that successfully has been applied for segmenting WMH on brain T2-FLAIR images (Rachmadi et al., 2019). Without any ground-truth segmentation, this algorithm produces a map which describes how much each voxel is irregular compared with an overall area. This map is usually called "irregularity map" (IM) or "irregularity age map" (IAM). The concept of this map was firstly suggested in the field of computer graphics to calculate pixel-wise "age" values indicating how weathered/damaged each pixel is compared to the overall texture pattern of an image (Bellini et al., 2016). Rachmadi et al. (2019) then proposed a similar approach to calculate the irregularity level of WMH with respect to the "normal" tissue in T2-FLAIR brain MRI (Rachmadi et al., 2017, 2018b). As WMH highlight irregular intensities on T2-FLAIR MRI slices, IAM can be also used for WMH segmentation. Although performing better than some conventional machine learning algorithms, LOTS-IM still underperforms compared to state-of-the-art deep neural networks. This is mainly because IAM essentially indicates irregular regions, including artifacts, other pathological features and some gray matter regions, in addition to WMH. However, considering IAM depicts irregularities quite accurately and can be generated without a training process, we propose to use IAM as an auxiliary guidance map of WMH location for WMH segmentation.

Recently, the introduction of deep neural networks, the state-of-art machine learning approach, has remarkably increased performances of image segmentation and object detection tasks. Deep neural networks outperform conventional machine learning approaches in bio-medical imaging tasks as well as general image processing. For example, Ciresan et al. (2012) built a pixel-wise classification scheme that uses deep neural networks to identify neuronal membranes on electron microscope (EM) images (Ciresan et al., 2012). In another study, Ronneberger et al. proposed a new deep neural network architecture called U-Net for segmenting neuronal structures on EM images (Ronneberger et al., 2015).

In medical images' segmentation tasks, U-Net architecture and its modified versions have been massively popular due to the end-to-end segmentation architecture and high performance. For instance, a U-Net-based fully convolutional network was proposed to automatically detect and segment brain tumors using multi-modal MRI data (Dong et al., 2017). A 3D U-Net for segmenting the kidney structure in volumetric images produced good quality 3D segmentation results (Çiçek et al., 2016). UResNet, which is a combination of U-Net and a residual network, was proposed to differentiate WMH from stroke lesions (Guerrero et al., 2018). Zhang Y. et al. (2018) trained a randomly initialized U-Net for WMH segmentation and improved the segmentation accuracy by post-processing the network's results (Zhang Y. et al., 2018).

While there have been many studies showing that U-Net performs well in image segmentation, it has one shortcoming that is long training time due to its high complexity (Briot et al., 2018; Zhang C. et al., 2018). To ameliorate this problem, Karargyros et al. suggested the application of regional maps as an additional input, for segmenting anomalies on CT images, and named their architecture Saliency U-Net (Karargyros and Syeda-Mahmood, 2018). They pointed out that extraction of relevant features from images unnecessarily demands very complex deep neural network architectures. Thus, despite neural networks architecture with large number of layers being able to extract more appropriate features from raw image data, it often accompanies a long training time and causes overfitting. Saliency U-Net has regional maps and raw images as inputs, and separately learns features from each data. The additional features from regional maps add spatial information to the U-Net, which successful delineates anomalies better than the original U-Net with less number of parameters (Karargyros and Syeda-Mahmood, 2018).

Another way to improve the segmentation performance of deep neural networks is through the recognition of the multi-scale context image information. Multi-scale learning is important particularly for detection/segmentation of objects with variable sizes and shapes. A dilated convolution layer was proposed to make deep neural networks learn multi-scale context better (Yu et al., 2017). Using dilated convolution layers, an architecture can learn larger receptive fields without significant increase in the number of parameters. Previous studies have reported improvements using dilated convolution layers in medical image processing tasks (Lopez and Ventura, 2017; Moeskops et al., 2017).

In this paper, we propose to use IAM as an additional input data to train a U-Net neural network architecture for WMH segmentation, owed to the fact that LOTS-IM can easily produce IAM without the need for training using manually marked WMH ground-truth data. U-Net architecture is selected as a base model for our experiments as it has shown the best learning performance using IAM (Rachmadi et al., 2018a). To address the incorporation of IAM to U-Net for WMH segmentation, we propose feed-forwarding IAM as regional map to a Saliency U-Net architecture. We also propose combining Saliency U-Net with dilated convolution to learn multi-scale context from both T2-FLAIR MRI and IAM data, in a scheme we name Dilated Saliency U-Net. We compare the original U-Net's performance with the performances of Saliency U-Net and Dilated Saliency U-Net on WMH segmentation.

Consequently, the contributions of our work can be summarized as follows:

- Proposing the use of IAM as an auxiliary input for WMH segmentation. T2-FLAIR MRI and IAM complement each other when they both are used as input to the neural network, addressing challenging cases especially those with few small WMH.

- Integration of Saliency U-Net and dilated convolution for WMH segmentation; which showed more detailed boundary delineation of large WMH. It also attained the best Dice coefficient score compared to our other experimental models.

2. MATERIALS AND METHODS

2.1. Dataset

MRI can produce different types of images to display normal tissues and different types of clinical abnormalities. It is desirable to choose suitable image types considering the properties of biomarkers or diseases targeted in the segmentation task. T2-weighted is one of the MRI sequences that emphasizes fluids as bright intensities. The bright intensity of fluids makes WMH difficult to identify in this MRI modality because WMH are also bright on T2-weighted. T2-fluid attenuated inversion recovery (T2-FLAIR) removes cerebrospinal fluid (CSF) signal from the T2-weighted sequence, increasing the contrast between WMH and other brain tissues. Therefore, we have chosen T2-FLAIR MRI as the main source of image data for our experiments.

We obtained T2-FLAIR MRI sequences from the public dataset *the Alzheimer's Disease Neuroimaging Initiative* (ADNI)[1] which was initially launched by Mueller et al. (2005). This study has mainly aimed to examine combinations of biomarkers, MRI sequences, positron emission tomography (PET) and clinical-neuropsychological assessments in order to diagnose the progression of mild cognitive impairment (MCI) and early AD. From the whole ADNI database, we randomly selected 60 MRI scans collected for three consecutive years from 20 subjects with different degrees of cognitive impairment in order to evaluate the applicability of our proposed scheme not only for cross-sectional studies but also for longitudinal analyses of WMH. Each MRI scan has dimensions of 256 × 256 × 35. We describe how train and test dataset are composed in section 2.8.

Ground truth masks were semi-automatically produced by an experienced image analyst using a thresholding algorithm combined with region-growing in the Object Extractor tool of Analyze^{TM} software. This semi-automatic WMH segmentation used the T2-FLAIR images. Intracranial volume (ICV) and CSF masks were generated automatically using optiBET (Lutkenhoff et al., 2014), and a multispectral algorithm developed in-house (Hernández et al., 2015), respectively. Full details and binary WMH reference masks can be downloaded from the University of Edinburgh DataShare repository[2].

2.2. Irregularity Age Map (IAM)

As described in section 1, the concept of IAM was proposed with the development of the LOTS-IM algorithm and its application to the task of WMH segmentation (Rachmadi et al., 2017, 2018b, 2019). This algorithm was inspired by the concept of "age map" proposed by Bellini et al. while calculating the level of weathering or damage of pixels compared to the overall texture pattern on natural images (Bellini et al., 2016). Rachmadi et al. adopted this principle to compute the degree of irregularity in brain tissue from T2-FLAIR MRI.

In this study, the GPU-powered LOTS-IM algorithm (Rachmadi et al., 2019)[3] was used to generate IAM from all scans. The steps of the LOTS-IM algorithm are as follows. Source and target patches are extracted from the MRI slices with four different sizes (i.e., 1 × 1, 2 × 2, 4 × 4, and 8 × 8) to capture different details in the brain tissues (Rachmadi et al., 2017). All grid fragments consisting of $n \times n$ sized patches are regarded as *source patches*. On the other hand, *target patches* are picked at random locations within the brain. Thus, non-brain target patches, located within the CSF mask or outside the ICV mask, are excluded from computation. Then, the difference between each source patch and one target patch on the same slice is calculated by Equation 1;

$$difference = \theta \cdot |max(s-t)| + (1-\theta) \cdot |mean(s-t)| \quad (1)$$

where s and t mean source patch and target patch, respectively, also θ was set to 0.5 (Rachmadi et al., 2018b). After difference values between a source patch and all target patches are calculated, the 100 largest difference values are averaged to become the *age value* of the corresponding source patch (Rachmadi et al., 2017). The rationale is that the average of the 100 largest difference values produced by an "irregular" source patch is still comparably higher than the one produced by a "normal" source patch (Rachmadi et al., 2017, 2018b). Furthermore, the age value is computed only for source patches within the brain to reduce the computational complexity. All age maps from four different patch sizes are, then, normalized to have normalized age values between 0 and 1; and each of them is up-sampled into its original image size and smoothed by a Gaussian filter. The final age map is produced by blending these four age maps using the Equation 2;

$$Final\ age\ map = \alpha \cdot AM_1 + \beta \cdot AM_2 + \gamma \cdot AM_4 + \delta \cdot AM_8 \quad (2)$$

where AM_x means the age map of $x \times x$ sized patches and $\alpha + \beta + \gamma + \delta = 1$. In this study, $\alpha = 0.65$, $\beta = 0.2$, $\gamma = 0.1$ and $\delta = 0.05$ (Rachmadi et al., 2019). Finally, the final age map is penalized by multiplying the original T2-FLAIR image slice to reflect only the high intensities of WMH, and globally normalized from 0 to 1 over all brain slices. The overall steps are schematically illustrated in **Figure 1**.

Though regarded as WMH segmentation map in the original studies, IAM essentially calculates the probability of each voxel to constitute an irregularity of the "normal" tissue. This irregular pattern includes not only WMH but more features such as artifacts, T2-FLAIR hyperintensities of other nature, as well as sections of the cortex that could be hyperintense. To compensate these flaws and take advantage of its usefulness, we developed a new scheme that uses IAM as an auxiliary guidance map for training deep neural networks rather than using it for producing the final WMH segmentation.

[1]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[2]https://datashare.is.ed.ac.uk/handle/10283/2214

[3]https://github.com/febrianrachmadi/lots-iam-gpu

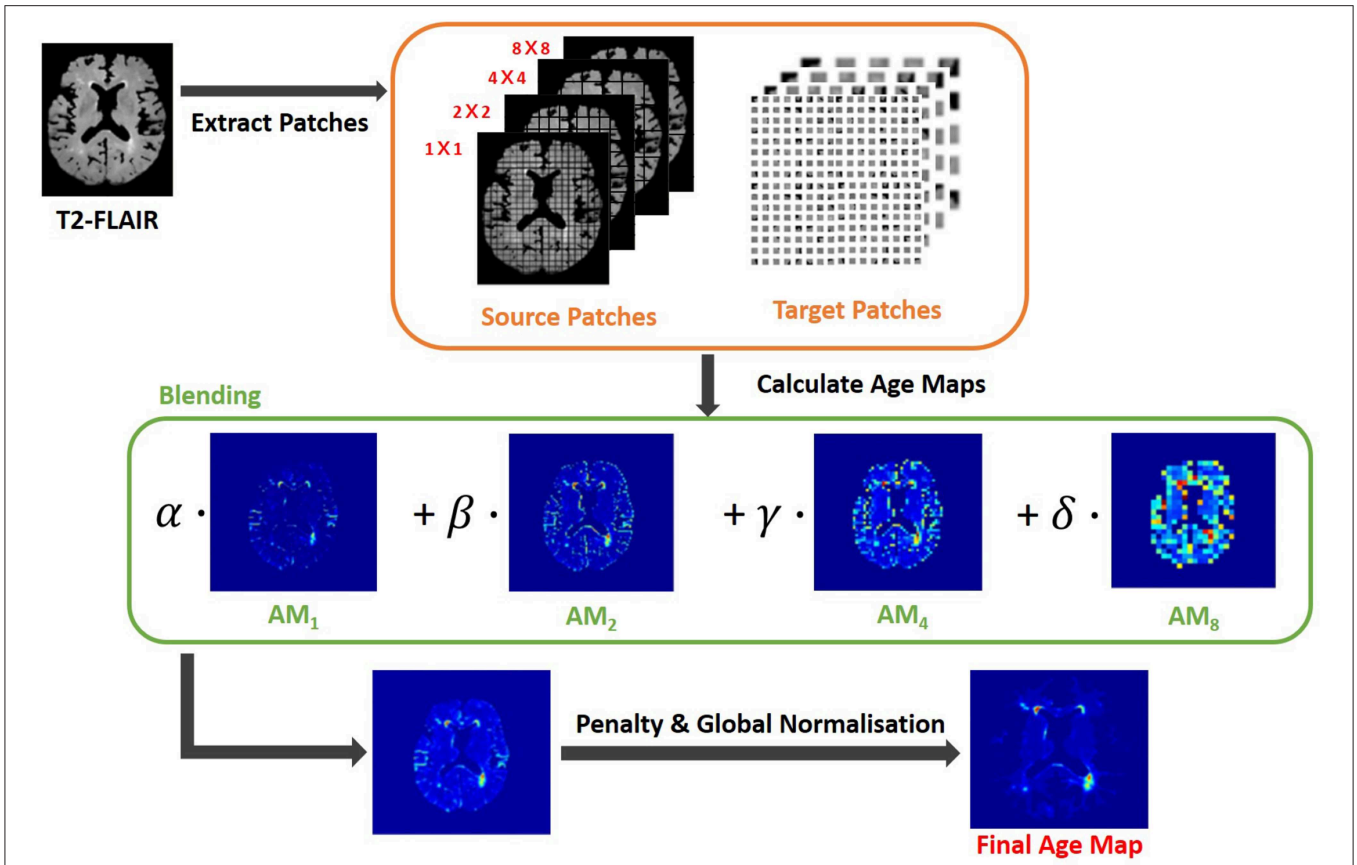

FIGURE 1 | Flow chart illustrating the LOTS-IM algorithm proposed by Rachmadi et al. (2019) applied to WMH segmentation. This study uses the final map generated by this algorithm, and refers to it as "IAM data".

2.3. U-Net

Since U-Net architecture was firstly presented (Ronneberger et al., 2015), various image segmentation studies have used this architecture due to its competitive performance regardless of the targeted object types. Different to the natural image segmentation, bio-medical image segmentation involves a more challenging circumstance as lack of data for the training process is a common problem. U-Net deals with this challenge with dense prediction of the input image using up-sampling layers that produce equal-sized input and output. This approach was drew by fully convolutional networks (Long et al., 2015).

U-Net is comprised of two parts, the encoding part where feature maps are down-sampled by max-pooling layers and the decoding part where the reduced size of feature maps are up-sampled to the original size. It retains the localization accuracy with the contracting path, which concatenates the feature maps stored in the encoding part with the decoding part. These kept high resolution features help to restore the details of localization removed by max-pooling layer, when the feature maps are up-sampled in the decoding part. The architecture is depicted in **Figure 3A**.

A drawback of U-Net is its large number of parameters. To restore the high resolution localization, the network should increase the number of feature channels in the decoding part. Training time and memory usage are proportional to the number of parameters. So training a U-Net architecture is constrained by its high consumption of time and memory. Moreover, the complexity of the (neural) network often induces the problem of overfitting.

2.4. Saliency U-Net

Saliency U-Net was first introduced to detect anomalies in medical images using a combination of raw (medical) images and simple regional maps (Karargyros and Syeda-Mahmood, 2018). Saliency U-Net performed better than U-Net while using less number of parameters. An architecture with less number of parameters is preferable as it is easier and faster to be trained. Karargyros and Syeda-Mahmood showed that convolution layers are not needed to extract more relevant features from raw images if auxiliary information from regional map is given as input. The Saliency U-Net architecture has two branches of layers in the encoding part (**Figure 3B**). Each branch extracts features from raw image and regional map independently, and the extracted features are fused before the decoding part.

Segmentation results from Saliency U-Net in the original study (Karargyros and Syeda-Mahmood, 2018) showed more

precise localization and better performance than the original U-Net, which contained a larger number of convolutional layers. Therefore, for WMH segmentation, we propose to use Saliency U-Net taking T2-FLAIR as raw input image and IAM as regional map.

2.5. Dilated Convolution

One common issue for image segmentation via deep neural networks is caused by the reduced size of the feature maps in the pooling layer introduced to capture global contextual information. While pooling layers are useful to get rid of some redundancies in feature maps, the lower size of feature maps after the last pooling layer also causes loses of some of its original details/information, decreasing the segmentation performance where the targeted regions are not spatially prevalent (Yu et al., 2017; Hamaguchi et al., 2018).

Dilated convolution solved this problem by calculating a convolution over a larger region without reducing the resolution (Yu and Koltun, 2015). The dilated convolution layer enlarges a receptive field including k skips between each input pixel. k is called *dilation factor*. In numerical form, a dilated convolution layer with a dilation factor k and a $n \times n$ filter is formulated as follows:

$$F(r,c) = \sum_{i=-n}^{i=n} \sum_{j=-n}^{j=n} W(i,j) I(r+ki, c+kj) \tag{3}$$

Figures 2A–C show examples of dilated convolution filters with dilation factors 1 to 3.

The additional advantage of dilated convolution is to widen the receptive field without increasing the number of parameters. Large receptive fields learn the global context by covering a wider area over the input feature map, but bring a memory leak and time consumption out for a growing number of parameters. Dilation can expand the receptive field of the convolution layer as much as skipped pixels without extra parameters. For instance, as shown in **Figures 2A,C**, the filter with dilation factor 3 has 7×7 sized receptive field, while the filter with dilation factor 1 has 3×3 sized receptive field.

In this study, we propose the incorporation of dilated convolution to Saliency U-Net for WMH segmentation. Since the size of WMH is variable, it is necessary to recognize different sizes of spatial contexts for more accurate delineation of WMH. We believe that dilated convolutions can manage the variable size of WMH from different sizes of receptive field.

2.6. Our Experimental Models

We examined three different U-Net models for which its original architecture was trained using input data with different modalities: T2-FLAIR (model 1), IAM (model 2), and both (model 3). To feed both T2-FLAIR and IAM together, we integrated T2-FLAIR and IAM as a two-channel input. As mentioned in section 2.3, U-Net architecture has encoding and decoding parts. In the encoding part, input images or feature maps are down-sampled by max-pooling layers to obtain relevant features for WMH segmentation. Then, in the decoding part, reduced feature maps are up-sampled again by up-sampling layers to acquire the original size in the final segmentation map. Max-pooling and Up-sampling layers are followed by two CONV blocks (yellow blocks in **Figure 3**). The CONV block contains a convolution layer, an activation layer and a batch normalization layer. Batch normalization allows to train neural networks with less careful initialization and higher learning rate by performing normalization at every batch (Ioffe and Szegedy, 2015). All activation layers except the last one are ReLU (Nair and Hinton, 2010), but the last activation layer calculates the categorical cross-entropy to yield a probability map for each label.

In addition, we trained Saliency U-Net and Dilated Saliency U-Net by feed forwarding both T2-FLAIR and IAM separately. In this way, we assume that IAM works as a simple regional map which provides localization information of WMH rather than just being a different image channel. While the U-Net architecture has one branch of the encoding part, Saliency U-Net encoding part consists of two branches that learn raw images and regional maps individually. Furthermore, we applied dilation factors of 1, 2, 4 and 2 to the first four convolutional layers of Saliency U-Net to form the Dilated Saliency U-Net. The architectures of U-Net, Saliency U-Net and Dilated Saliency U-Net can be seen in **Figure 3**.

Performance of these models are compared to each other in section 3. We additionally conducted experiments on the original U-Net models trained only with T2-FLAIR and only with IAM in order to see how using both T2-FLAIR and IAM as inputs affects learning WMH segmentation. Our five experimental models are listed in **Table 1**.

2.7. Preprocessing

In machine learning, data preprocessing is needed to standardize the data into a comparable range. It is especially important when we deal with MRI data whose intensity is not in a fixed range. Differences in the intensity range are caused by differences in MRI acquisition protocols, scanner models, calibration settings, etc. (Shah et al., 2011).

For this reason, we normalized the intensity of the brain tissue voxels in our train and test data. The image intensity of the majority of non-brain tissue voxels of an MRI slice is zero or near-zero, although few non-brain voxels can have peak intensity values above the intensity range of the brain tissue. Thus, normalizing intensities from all voxels together can bias the intensity values toward zero and reduce the effect of WMH on brain tissue voxels. Brain tissue voxels were filtered using CSF and the intracranial volume (ICV) masks as follows:

$$Brain\ Tissue\ Region\ =\ MRI\ scan \cap (\neg\, CSF \cap\ ICV\,) \tag{4}$$

We normalized the brain tissue voxels on each slice into a distribution with zero-mean and unit variance by subtracting the mean value from each voxel value and dividing the result by the standard deviation.

Although WMH segmentation can be regarded as the binary classification of voxels, we re-labeled the ground-truth data assigning voxels one of the three following labels: non-brain, non-WMH brain tissue and WMH. However, when evaluating the segmentation results, we considered both non-brain and

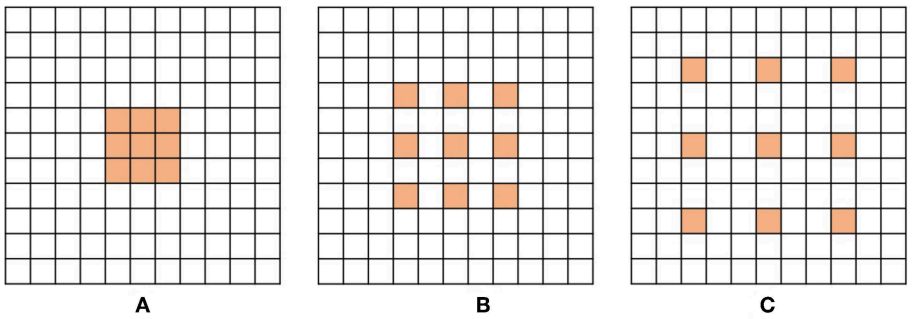

FIGURE 2 | Examples of dilated convolution filter with 3 × 3 size. **(A)** Dilation factor = 1, **(B)** Dilation factor = 2 and **(C)** Dilation factor = 3.

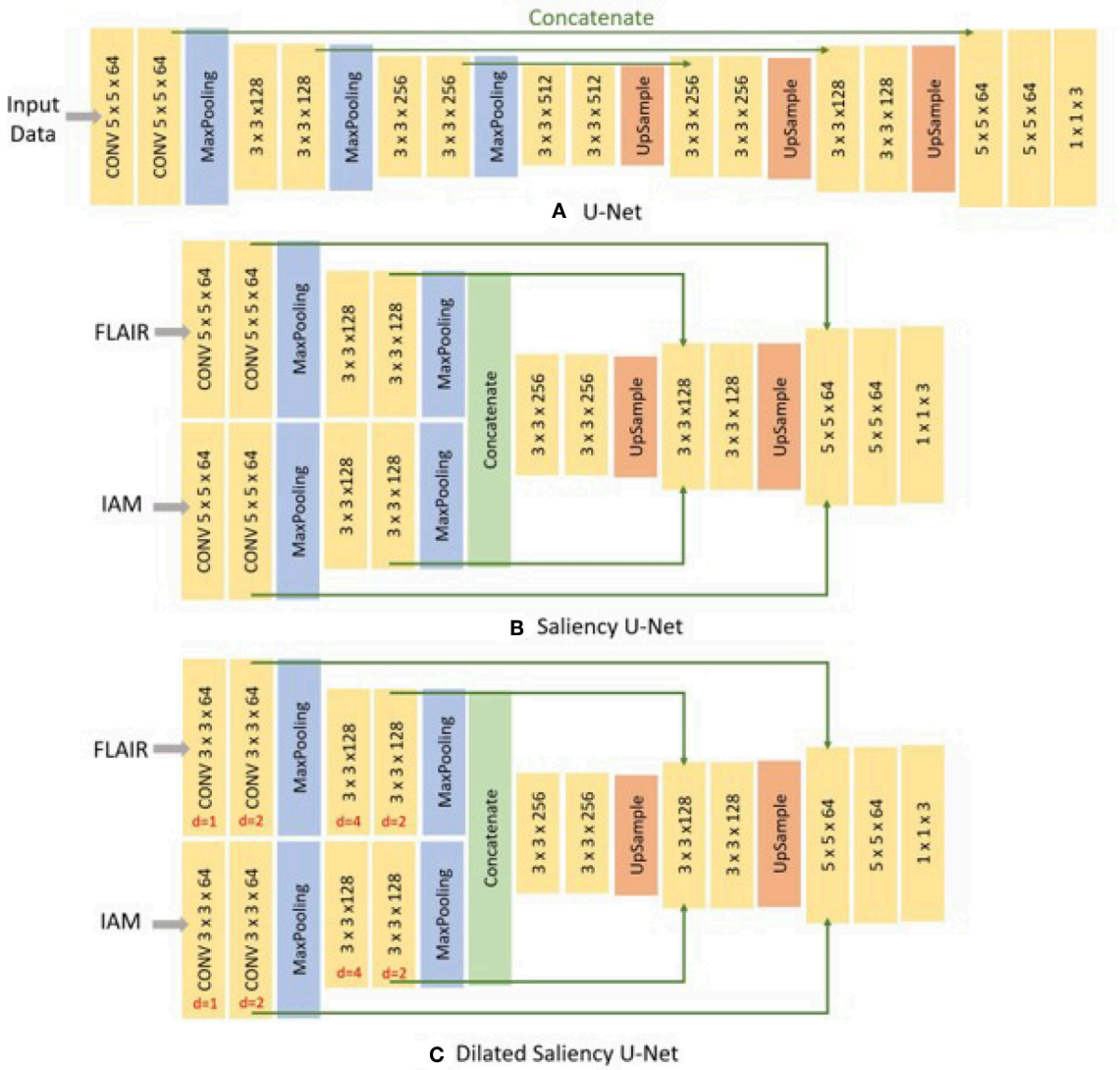

FIGURE 3 | Architecture of three different networks used in this study. **(A)** the original U-Net, **(B)** Saliency U-Net, and **(C)** Dilated Saliency U-Net. Three numbers of CONV block (yellow block) represents *filter size* × *filter size* × *filter channels*. For the Dilated Saliency U-Net model, red numbers mean a dilation factor for the convolution layer in each CONV block.

TABLE 1 | Dice Similarity Coefficient (DSC), sensitivity, positive predictive value (PPV), training time and number of parameters for our five experimental models.

Model	DSC	Sensitivity	PPV	Training time	# Parameters
U-Net(FLAIR)	0.5440	0.4594	0.6275	1 h 52 m 55 s	7,859,715
U-Net(IAM)	0.5274	0.4179	**0.6769**	1 h 53 m 52 s	7,859,715
U-Net(F+I)	0.5281	**0.4902**	0.6268	1 h 24 m 22 s	7,861,315
Saliency U-Net(F+I)	*0.5535*	0.4730	0.6034	1 h 30 m 1 s	2,756,803
Dilated Saliency U-Net(F+I)	**0.5588**	*0.4747*	*0.6374*	1 h 4 m 18 s	2,623,683

Values in bold are the highest scores and in italic the second highest. In the brackets after the model names, the input data type is specified. "FLAIR" is equivalent to T2-FLAIR and "F+I" refers to taking both T2-FLAIR and IAM as input.

non-WMH brain tissue labels as non-WMH labels to calculate sensitivity and Dice similarity coefficient which are metrics for the binary classification. **Figure 4** shows the example of a T2-FLAIR slice, the same slice after preprocessing and normalization, and the ground-truth slice.

2.8. Training and Testing Setup

For training, 30 MRI scans of the ADNI dataset described in section 2.1 were randomly selected. These 30 MRI scans were collected from 10 subjects for three consecutive years. We trained our networks with image patches generated from these MRI scans, not slices, to increase the amount of training data. If we train our models using slice images, the amount of training data is only $35 \times 30 = 1050$ slices, which is not ideal for training a deep neural network architecture. Instead, by extracting 64×64 sized patches from each image slice, we could have 30,000 patches for training data.

For testing, we used the rest 30 scans of the ADNI sample, which are not used during training. These scans were also obtained from another 10 subjects for three consecutive years. The testing dataset was comprised of image slices without patch extraction. Slice image data is necessary to analyse the results from our models according to the distributions or volumes of WMH. Our testing dataset holds 1050 of 256×256 image slices in total as each scan contains 35 slices.

All experimental models were trained using the same network configuration. We set learning rate to $1e^{-5}$ and batch size to 16. As an optimization method, we selected the Adam optimization algorithm (Kingma and Ba, 2014), although the original U-Net scheme used the stochastic gradient descent (SGD) optimizer. This is because the Adam optimizer can handle sparse gradients. It is highly possible that our training data produce sparse gradients as non-brain voxels, which are the majority, have zero intensity. We applied the Adam optimizer accordingly, considering this data property.

3. RESULTS

In this section, we present how experiments were conducted, and analyse and compare the experimental results.

3.1. Evaluation Metrics

We use sensitivity, positive predictive value (PPV) and Dice similarity coefficient (DSC) to evaluate the models. Sensitivity measures the rate of true positives as below:

$$Sensitivity = \frac{TP}{TP + FN} \tag{5}$$

where *TP* means true positive, and *FN* means false negative. PPV also measures the rate of true positives but from the total of positive calls like below:

$$PPV = \frac{TP}{TP + FP} \tag{6}$$

where FP refers to false positive. DSC is a statistic method to compare the similarity between two samples of discrete values (Dice, 1945). It is one of the most common evaluation metrics in image segmentation. The formula is as follow:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{7}$$

where *TP* and *FN* are as per Equation 5 and *FP* means false positive. DSC is interpreted as the overlapping ratio to the whole area of prediction and target objects, while sensitivity measures the correctly predicted region of the target object. If the prediction includes not only true positives but also wrong segmentation results (false positives), the DSC score can be low despite the high sensitivity.

3.2. The Effects of IAM as an Auxiliary Input Data

Table 1 shows overall performances of our five experimental models. The adoption of IAM as an auxiliary input data for U-Net [i.e., U-Net(F+I)] improved sensitivity to 0.4902 but had lower DSC score than the model that used only the T2-FLAIR image as input. On the other hand, Saliency U-Net(F+I) improved the DSC scores achieved by U-Net to 0.5535 while Dilated Saliency U-Net(F+I) achieved the best DSC score of 0.5588. Dilated Saliency U-Net(F+I) yielded the second best sensitivity rate after U-Net trained with T2-FLAIR and IAM [i.e., U-Net(F+I)]. U-Net(IAM) achieved the best PPV value of our five models and Dilated Saliency U-Net(F+I) achieved the second highest value of PPV. From these results, we can see that the three models trained with T2-FLAIR and IAM particularly increased the sensitivity performance of the network architectures.

Saliency and Dilated Saliency U-Net included considerably less parameters than the three U-Net models. As shown in **Table 1**, Saliency and Dilated Saliency U-Net have more than three times less parameters and slightly shorter training time than the original U-Net while having better if not similar performance on WMH segmentation.

With regards to training time, although feeding both T2-FLAIR and IAM together into U-Net involved the calculation of more parameters due to the two-channel input, the training time for this model was shorter than that of U-Net(FLAIR) and U-Net(IAM). In deep learning studies, visual attention, which gives larger weight on the region of interest, speeds up learning by leading the model to concentrate on the relevant regions.

This has been experimentally demonstrated in previous studies (Choi et al., 2017; Najibi et al., 2018). In our case, IAM confers the visual attention effect to the network architecture. Despite having fewer parameters, Saliency U-Net took longer time to train than U-Net(F+I). Feed-forward and back-propagation proceed separately in each encoding part. Dilated Saliency U-Net significantly decreased the training time compared to the other models by skipping voxels that reduce the computational complexity, when calculating the convolution.

Figure 5 presents training and validation losses for our five models. Same color lines correspond to the same model. Solid and dashed lines represent training loss and validation loss each. For all models, both training and validation losses properly converged. Thus, our models are not overfitted on the training data.

We also evaluated whether the median and the distribution of DSC scores throughout the testing set differed significantly between the five models evaluated. We conducted two tests: (1) the Wilcoxon ranksum, as implemented by the function `ranksum` in MATLAB, to evaluate whether the medians of the DSC scores from each model across the testing dataset were significantly different between each other; and (2) the Kruskal-Wallis test, as implemented by the MATLAB function `kruskalwallis`, to evaluate whether the distributions of these DSC values were statistically significantly different between the models. Neither the medians nor the DSC distributions obtained by these five models significantly differed. The result of the Kruskal-Wallis test is shown in **Table 2**. The *p-value* obtained from the ANalysis Of VAriance (ANOVA) of the DSC distributions from the five models across all cases is 0.7786, indicating that the results of these five models did not differ significantly from each other in terms of the distribution of DSC across the testing set. This emphasizes that Dilated Saliency U-Net model can produce similar level of performance as the original U-Net models even with less number of parameters and shorter training time. **Figure 6** also illustrates that the DSC scores obtained from applying our models are similarly distributed to each other.

Figure 7 visualizes the examples of WMH segmentation results by our experimental models. In most cases, the use of two data sources (i.e., IAM and T2-FLAIR images) in training the

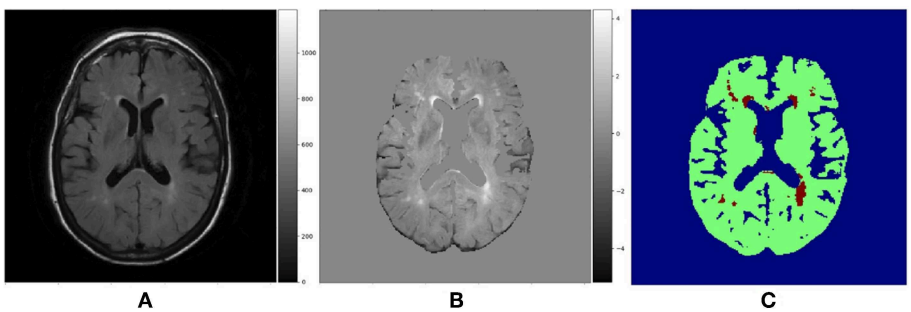

FIGURE 4 | (A) Raw T2-FLAIR image, **(B)** T2-FLAIR input after preprocessing and normalization, **(C)** Ground truth data with three labels. Blue region is non-brain area, green region is non-WMH brain tissues and red region is WMH.

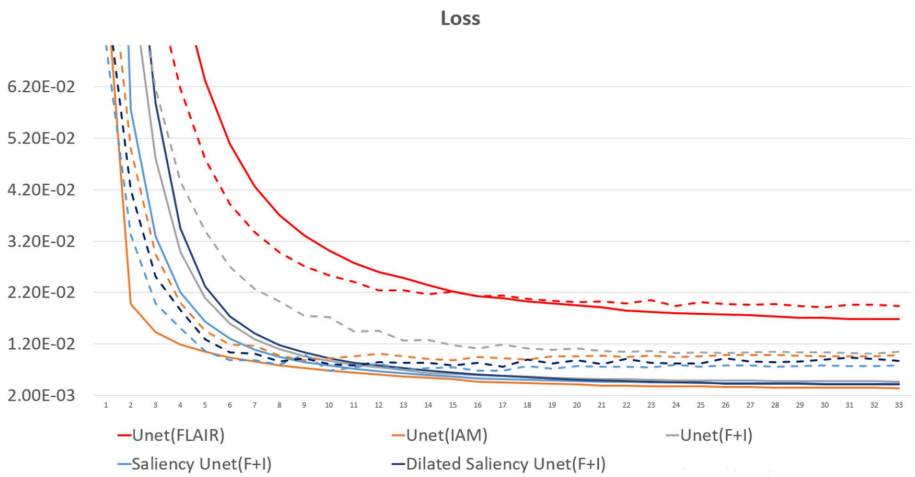

FIGURE 5 | Loss graph of our five models. While solid lines indicate training loss, dashed lines represent validation loss.

network complements each other's effect detecting tricky WMH regions. Depending on the contrast/size of WMH or the quality of IAM, there are some cases in which WMH are distinguishable on IAM but unclear in T2-FLAIR and vice versa. For example, if WMH clusters are too small, it is hard to differentiate them on T2-FLAIR, but they are better observable on IAM, where WMH and normal brain tissue regions have better contrast. On the other hand, in the presence of other irregular patterns such as extremely low intensities of brain irregularities around WMH, T2-FLAIR can indicate WMH clearly than IAM. In **Figure 7A**, U-Net(FLAIR) produced better WMH segmentation result than U-Net(IAM) due to the poor quality of IAM. Conversely, U-Net(FLAIR) could not detect WMH well due to unclear intensity contrast on T2-FLAIR while U-Net(IAM) could segment these WMH regions as IAM enhanced them as anomalies (**Figure 7B**). Furthermore, incorporating both T2-FLAIR and IAM together as input data produced better WMH segmentation in general (5–7th columns from left to right of **Figure 7**).

TABLE 2 | ANOVA table for our five models.

Source	SS	df	MS	*F-value*	*p-value*
Models	3334.7	4	833.68	1.77	0.7786

SS refers to the sum of squares. df and MS mean degrees of freedom and mean squares, respectively.

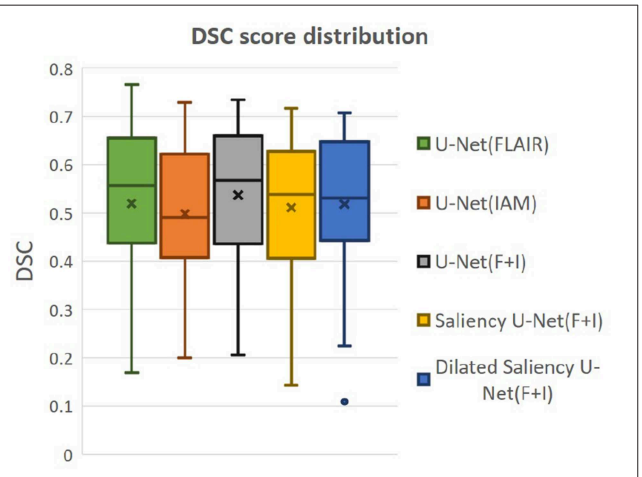

FIGURE 6 | Distributions of DSC score by our five models. "x" and bar at the middle of box indicate mean and median each. Bottom and top of each box mean the first and third quartile.

3.3. WMH Volume Analysis

In this experiment, we evaluate our models based on the WMH volumes of the MRI scan (i.e., WMH burden) to examine the influence of WMH burden on the performance of WMH segmentation. The WMH volume of each MRI scan is calculated by multiplying the number of WMH voxels by the voxel size. We grouped MRI scans into three groups according to the range of WMH volume. **Table 3** shows the range of WMH volume used as criteria for forming the groups, and the number of scans included in each group. **Figure 8A** shows the lack of ambiguity or overlap in the classification of the MRI scans in each group.

Figure 8B plots the DSC scores yielded by the MRI scans in the different WMH volume groups by our five experimental

TABLE 3 | Criteria sorting MRI scans according to WMH voxel volume.

Group	Range of WMH volume (mm^3)	# Scans
Large	$10,000 \leq$ WMH Vol	6
Medium	$4,000 \leq$ WMH Vol $< 10,000$	10
Small	$1 \leq$ WMH vol $< 4,000$	14

"# Scans" means the number of included MRI scans. Most of scans are included in Small and Medium groups.

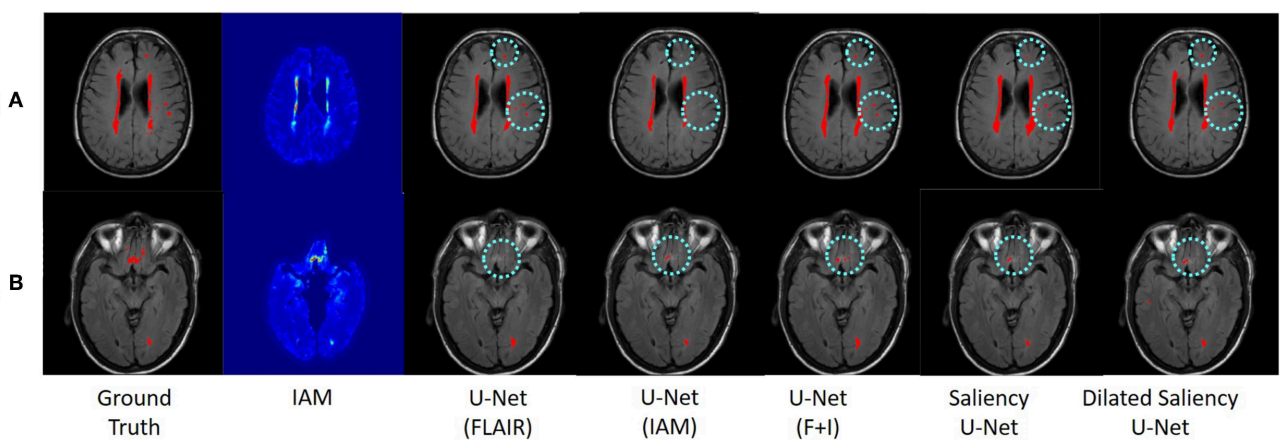

FIGURE 7 | Examples of WMH segmentation results by our experimental models. Cyan circles indicate WMH detected only by one of the original U-Net models, i.e., U-Net(FLAIR) or U-Net(IAM). Row **(A)** shows a case where WMH is distinguishable exclusively in the T2-FLAIR image, while row **(B)** shows a case where IAM highlights WMH clearly. By training networks using T2-FLAIR and IAM, both WMH regions are detected.

models. Please, note that the DSC scores referred in this section correspond to the evaluation of the WMH segmentation results in each MRI scan, not per slice which are used for overall performance evaluation in section 3.2 **Table 1**. Hence scans of the Large group might have several small WMH rather than one large region with confluent WMH.

All models tested in this study showed high median values of DSC scores in the Medium group, for which all models performed better than the other groups. In the Large group, U-Net(FLAIR) and U-Net(F+I) models performed similarly well, while U-Net(IAM) performed worst compared with the rest of the models. Mean, median and standard deviation (std.) values of DSC score distribution in each group are shown in **Table 4**. Overall, the performance of the models for scans with Small and Medium WMH burden was quite similar (see also **Figure 8B**). However, large variations in DSC scores were observed among the scans of the Small group, especially for the U-Net(FLAIR) model.

3.4. Longitudinal Evaluation

In the Longitudinal evaluation test we addressed the capacity of our five models in predicting WMH in subsequent years after being trained only using the first year samples. Hence, the training set was formed by the first year samples while the testing set was composed by the second and third year samples. **Table 5** shows the mean DSC score for each sample. In this evaluation, U-Net(IAM) and Saliency U-Net performed slightly better than the other three models, partly owed to IAM which could provide information to predict WMH occurrence. As expected, all our models predicted better WMH in the second year than in the third year.

3.5. U-Net vs. Saliency U-Net

In order to evaluate the effectiveness of the Saliency U-Net architecture, we compared the original U-Net and Saliency U-Net models trained with T2-FLAIR and IAM. As shown in **Table 1**, Saliency U-Net yielded higher DSC score than U-Net(F+I) despite U-Net(F+I) having higher sensitivity value.

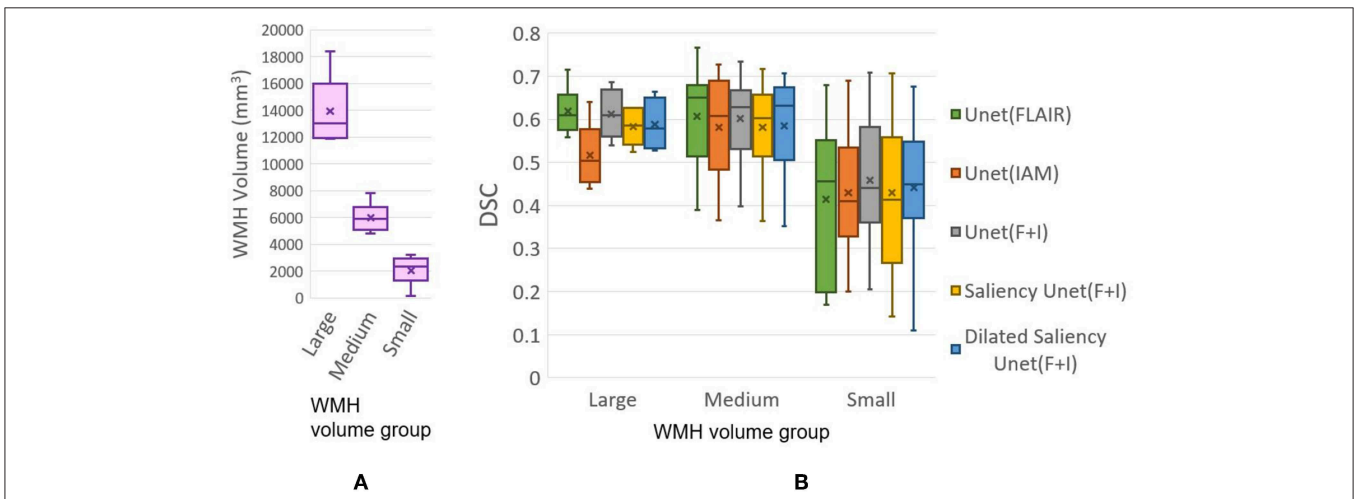

FIGURE 8 | (A) Distributions of data (MRI scan) grouped together based on WMH volume. **(B)** DSC distributions yielded by tested five models based on WMH volume. "x" and bar at the middle of box indicate mean and median each. Bottom and top of each box mean the first and third quartile.

TABLE 4 | Mean, median and standard deviation values of the distributions of DSC scores from our experimental models per WMH volume groups.

Model	DSC mean.			DSC median.			DSC std.		
	Large	Medium	Small	Large	Medium	Small	Large	Medium	Small
U-Net(FLAIR)	0.6184	0.6070	0.4147	0.6987	0.6499	0.4559	0.0524	0.1076	0.1746
U-Net(IAM)	0.5168	0.5817	0.4294	0.5036	0.6080	0.4106	0.0668	0.1111	0.1455
U-Net(F+I)	0.6124	0.6025	0.4580	0.6092	0.6276	0.4400	0.0548	0.0931	0.1460
Saliency U-Net(F+I)	0.5824	0.5812	0.4299	0.5853	0.6023	0.4134	0.0377	0.0956	0.1687
DSU-Net_1224	0.5722	0.5929	0.4003	0.5814	0.6286	0.3876	0.0592	0.0965	0.1733
DSU-Net_4221	0.5711	0.5768	0.4253	0.5776	0.6152	0.4250	0.0574	0.1097	0.1640
DSU-Net_1242	0.5882	0.5852	0.4407	0.5782	0.6320	0.4498	0.0536	0.1069	0.1558

Model name DSU-Net_abcd refers to Dilated Saliency U-Net model with dilation factors a, b, c, d in order from the first to the fourth convolution layers. These dilation factors are applied on convolution layers in the encoding part (i.e., before concatenating T2-FLAIR and IAM feature maps) of the CONV blocks, which consists of convolution, ReLU, and batch normalization layers. These different Dilated Saliency U-Net models are described in section 3.6. DSU-Net_1242 was used for the Dilated Saliency U-Net model evaluated in section 3.3.

Figure 9 shows that Saliency U-Net successfully eliminates some of the false positives observed in the segmentation result from U-Net(F+I).

We also investigated the change in Saliency U-Net's performance in relation to its complexity when the number of convolution layers increased/decreased. DSC score, training time and model complexity (i.e., the number of parameters) are compared in **Figure 10**. The rule for changing the Saliency U-Net complexity is to connect/disconnect the 2 CONV blocks that are attached/detached at both ends, through a "skip" connection. However, since the encoder part is a two-branch architecture, 6 CONV blocks are included at once increasing its complexity (i.e., 4 CONV blocks are added to the encoder part and 2 CONV block are added to the decoder part). Similar approach is done when decreasing the complexity, where 4 CONV blocks and 2 CONV blocks are dropped from the encoder and decoder, respectively. For clarity, our original Saliency U-Net model (i.e., evaluated in **Table 1** of section 3.2) contains 14 CONV blocks and each CONV block holds one convolution layer as shown in **Figure 3**.

As shown in **Figure 10**, adding more CONV blocks means increasing both number of parameters and training time significantly. Furthermore, using too many CONV blocks (i.e., Saliency U-Net with 26 CONV blocks) decreased the DSC score due to overfitting.

TABLE 5 | DSC score for longitudinal evaluation of our five models.

Model	2nd year	3rd year
U-Net(FLAIR)	0.6136	0.5878
U-Net(IAM)	**0.6270**	*0.6110*
U-Net(F+I)	0.6229	0.5823
Saliency U-Net	*0.6258*	**0.6119**
Dilated Saliency U-Net	0.6060	0.5881

*We evaluated these models using data from both second and third years. As per **Table 1**, values in bold are the highest scores and in italics are the second highest ones.*

3.6. Exploration of Dilated Saliency U-Net Architecture

In this experiment, we applied different dilation factors in Dilated Saliency U-Net, which captures multi-context information on image slices without having to change the number of parameters. As per **Figure 9**, which visually displays the segmentation results from Saliency U-Net, the boundary delineation is still poor for large WMH regions. Furthermore, we also can see in the same **Figure 9** that dilated convolutions help Saliency U-Net to reproduce the shape of WMH regions in more detail. Hence, it is important to know the influence of different dilated convolution configurations in Dilated Saliency U-Net for WMH segmentation.

In order to find the most appropriate dilation factors, we compared different sequences of dilation factors. **Figure 3C** shows the basic Dilated Saliency U-Net architecture used in this experiment. Only four dilation factors in the encoding part were altered while the rest of the parameters for the training schemes stayed the same. Yu and Koltun suggested to use a fixed filter size for all dilated convolution layers but exponential dilated factors (e.g., 2^0, 2^1, 2^2 ...) (Yu and Koltun, 2015). Therefore, we assessed "increasing", "decreasing" and "increasing & decreasing" dilation factor sequences with factor numbers of 1, 2, 2, 4 and fixed filter size of 3×3. Details of these configurations are presented in **Table 6**. From this table, we can appreciate that despite DSU-Net_4221 performed best in DSC score (0.5622), it recorded the lowest sensitivity score. The best sensitivity metric was produced by DSU-Net_1242 (0.4747), but it did not outperform DSU-Net_4221 in DSC score.

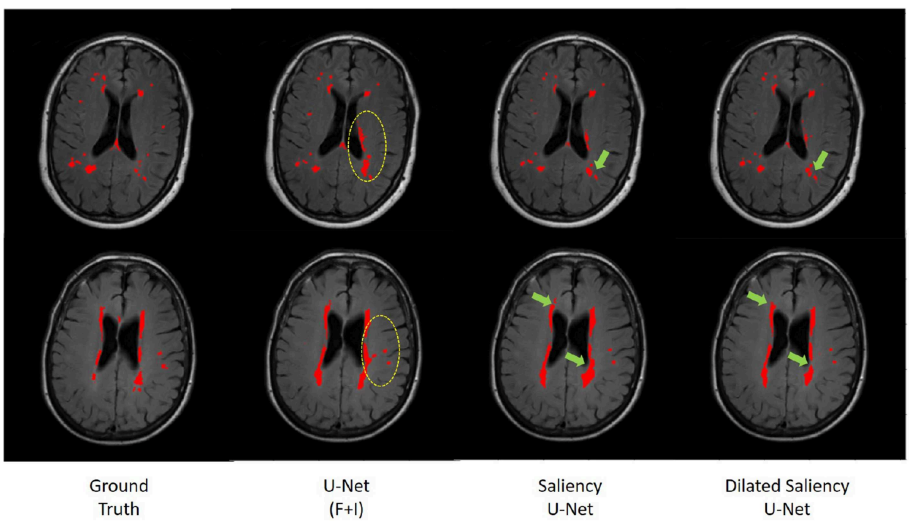

FIGURE 9 | Comparison of WMH segmentation results from U-Net(F+I), Saliency U-Net and Dilated Saliency U-Net. Yellow circles indicate false positive results by U-Net(F+I). These false positive results are eliminated in the results from Saliency and Dilated Saliency U-Net. Green arrows are pointing to locations where boundaries are segmented in more detail by Saliency and Dilated Saliency U-Net.

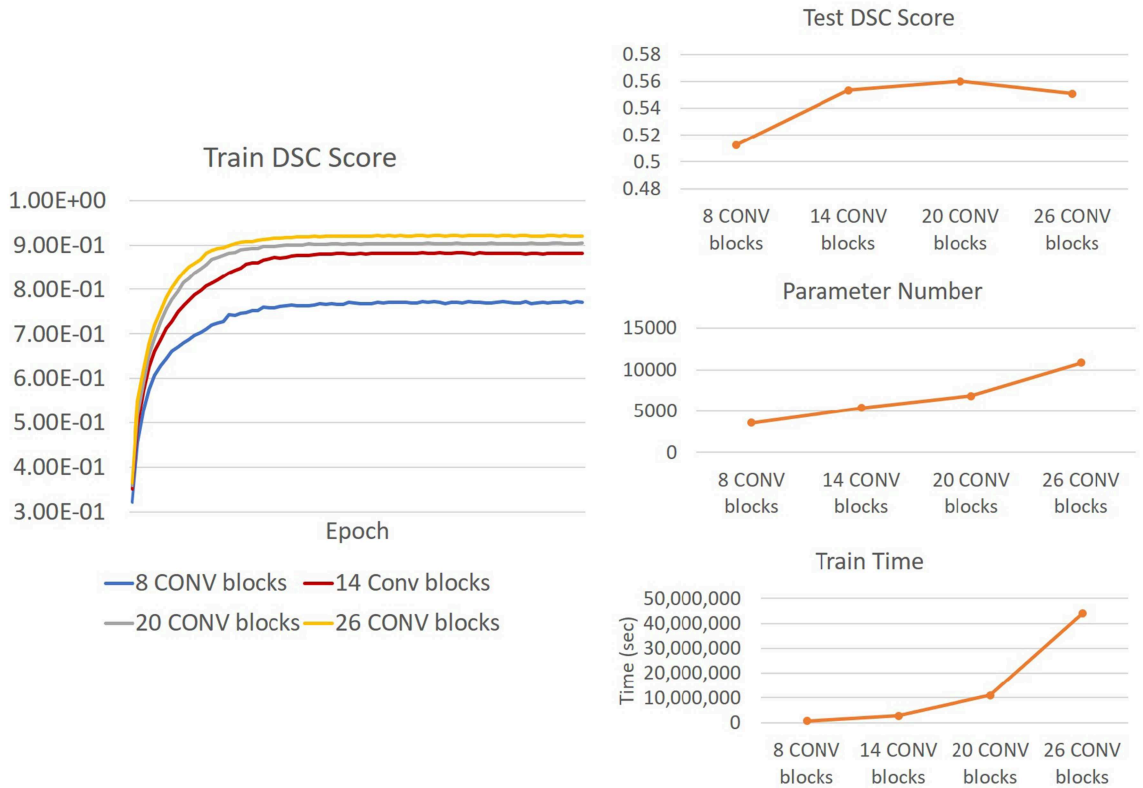

FIGURE 10 | (Right) Trends of DSC score, training time and number of parameters of Saliency U-Net when more convolution layers are changed. It is also shown Saliency U-Net with 26 CONV blocks performance in testing **(upper Right)** decreases although its training **(Left)** DSC score increase. This is caused by overfitting.

TABLE 6 | Encoder architecture of Dilated Saliency U-Net with different dilation factors and their performances.

Model	Encoder	DSC	Sensitivity
DSU-Net_1224 (Increasing)	CONV 3 × 3 × 64, d = 1 CONV 3 × 3 × 64, d = 2 Max Pooling CONV 3 × 3 × 128, d = 2 CONV 3 × 3 × 128, d = 4 Max Pooling	0.5304	0.4395
DSU-Net_4221 (Decreasing)	CONV 3 × 3 × 64, d = 4 CONV 3 × 3 × 64, d = 2 Max Pooling CONV 3 × 3 × 128, d = 2 CONV 3 × 3 × 128, d = 1 Max Pooling	**0.5622**	0.4381
DSU-Net_1242 (Increasing & decreasing)	CONV 3 × 3 × 64, d = 1 CONV 3 × 3 × 64, d = 2 Max Pooling CONV 3 × 3 × 128, d = 4 CONV 3 × 3 × 128, d = 2 Max Pooling	0.5588	**0.4747**

Three numbers in the CONV block stans for "filter size × filter size × filter number" and "d" means a dilation factor and its trend of dilation factor pattern is specified in the bracket. Values in bold are the highest scores.

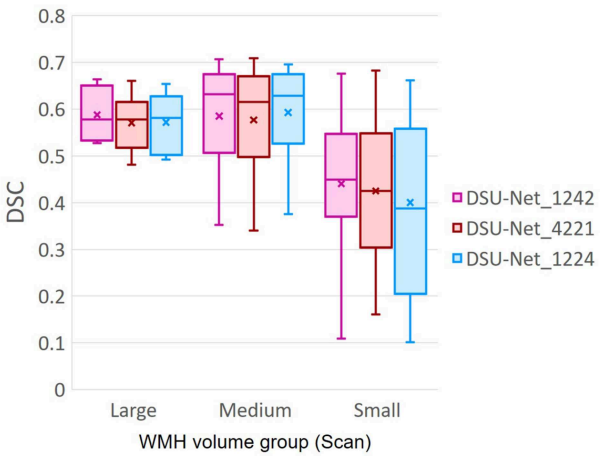

FIGURE 11 | DSC score of three different Dilated Saliency U-Net groups based on WMH volume in MRI scans. The group information is described in **Table 3**. "x" and bar at the middle of box indicate mean and median each. Bottom and top of each box means the first and third quartile.

Additionally, we investigated the influence of dilation factors in DSC score performance per WMH volume of MRI scans. Evaluation was conducted on the three groups previously described in **Table 3**. **Figure 11** shows that DSU-Net_1242

outperformed other models in every group. The report of mean, median and standard deviation of DSC score distribution in each group can be seen in **Table 4**.

4. DISCUSSION

In this study, we explored the use of IAM as an auxiliary data to train deep neural networks for WMH segmentation. IAM produces a probability map of each voxel to be considered a textural irregularity compared to other voxels considered "normal" (Rachmadi et al., 2019). While incorporating IAM as an auxiliary input data, we compared three deep neural network architectures to find the best architecture for the task, namely U-Net, Saliency U-Net and Dilated Saliency U-Net. It has been suggested that Saliency U-Net is adequate to learn medical image segmentation task with both a raw image and a pre-segmented regional map (Karargyros and Syeda-Mahmood, 2018). The original U-Net did not improve DSC score despite using both T2-FLAIR and IAM as input, but the DSC score from Saliency U-Net was superior to that from the original U-Net trained only with T2-FLAIR. This is because Saliency U-Net is able to learn the joint encoding of two different distributions: i.e., from T2-FLAIR and IAM. Saliency U-Net generated better results than U-Net despite having less parameters. We also found that Saliency U-Net had lower false positive rate compared to U-Net.

Dilated convolution can learn spatially multi-context by expanding the receptive field without increasing the number of parameters. We added dilation factors to the convolution layers in the encoding block of Saliency U-Net to improve WMH segmentation, especially due to the high variability in the WMH size. This new model is named "Dilated Saliency U-Net." Dilated convolution improved both DSC score and sensitivity with shorter training time. Dilated Saliency U-Net also yielded more accurate results in the presence of large WMH volumes and worked well in Medium and Small WMH volume MRI data groups which are more challenging. We identified that dilated convolution is effective when dilation factors are increased and decreased sequentially.

To our knowledge, this is the first attempt of successfully combining dilation, saliency and U-Net. We could reduce the complexity of a deep neural network architecture while increasing its performance through the integrated techniques and the use of IAM. Due to the trade-off between performance and training time, which is proportional to the model complexity, it is crucial to develop less complex CNN architectures without decreasing their performance.

Anomaly detection in the medical imaging field has been broadly studied (Quellec et al., 2016; Schlegl et al., 2017). One of its difficulties relies on the inconsistent shape and intensity of these anomalies. IAM helped the CNN scheme to overcome this problem by providing the localization and morphological information of irregular regions. We believe it is possible to generate IAM from different modalities of medical images. Thus, the application of IAM is highly expandable to detect different imaging bio-markers involving abnormal intensity values in other diseases.

AUTHOR CONTRIBUTIONS

YJ, MR, MV-H, and TK conceived and presented the idea. YJ and MR planned the experiments. YJ carried out the experiments. All authors provided critical feedback and analysis, and edited the manuscript.

FUNDING

Funds from the Indonesia Endowment Fund for Education (LPDP) of Ministry of Finance, Republic of Indonesia (MR) and Row Fogo Charitable Trust (Grant No. BRO-D.FID3668413) (MV-H) are gratefully acknowledged. This project is partially funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC) through the International Partnership Award BB/P025315/1 to MV-H. This study uses data from 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 and Development, LLC.; Johnson and Johnson Pharmaceutical Research and Development LLC.; Lumosity; Lundbeck; Merck and 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.

REFERENCES

Admiraal-Behloul, F., Van Den Heuvel, D., Olofsen, H., van Osch, M. J., van der Grond, J., Van Buchem, M., et al. (2005). Fully automatic segmentation of white matter hyperintensities in mr images of the elderly. *Neuroimage* 28, 607–617. doi: 10.1016/j.neuroimage.2005.06.061

Bellini, R., Kleiman, Y., and Cohen-Or, D. (2016). Time-varying weathering in texture space. *ACM Trans. Graph.* 35:141. doi: 10.1145/2897824.2925891

Bowles, C., Qin, C., Guerrero, R., Gunn, R., Hammers, A., Dickie, D. A., et al. (2017). Brain lesion segmentation through image synthesis and outlier detection. *NeuroImage Clin.* 16, 643–658. doi: 10.1016/j.nicl.2017.09.003

Briot, A., AI, G., Creteil, V., Viswanath, P., and Yogamani, S. (2018). "Analysis of efficient cnn design techniques for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Salt Lake City, UT), 663–672.

Choi, J., Lee, B.-J., and Zhang, B.-T. (2017). "Multi-focus attention network for efficient deep reinforcement learning," in *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, CA).

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 424–432.

Ciresan, D., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2012). "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Lake Tahoe: Curran Associates, Inc.), 2843–2851. Available online at: http://papers.nips.cc/paper/4741-deep-neural-networks-segment-neuronal-membranes-in-electron-microscopy-images.pdf

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302.

Dong, H., Yang, G., Liu, F., Mo, Y., and Guo, Y. (2017). "Automatic brain tumor detection and segmentation using u-net based fully convolutional networks," in *Annual Conference on Medical Image Understanding and Analysis* (Edinburgh, UK: Springer), 506–517.

Gootjes, L., Teipel, S., Zebuhr, Y., Schwarz, R., Leinsinger, G., Scheltens, P., et al. (2004). Regional distribution of white matter hyperintensities in vascular dementia, alzheimer's disease and healthy aging. *Dement. Geriat. Cogn. Disord.* 18, 180–188. doi: 10.1159/000079199

Guerrero, R., Qin, C., Oktay, O., Bowles, C., Chen, L., Joules, R., et al. (2018). White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. *NeuroImage Clin.* 17, 918–934. doi: 10.1016/j.nicl.2017.12.022

Hamaguchi, R., Fujita, A., Nemoto, K., Imaizumi, T., and Hikosaka, S. (2018). "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Lake Tahoe: IEEE), 1442–1450. doi: 10.1109/WACV.2018.00162

Hernández, M. d. C. V., Armitage, P. A., Thrippleton, M. J., Chappell, F., Sandeman, E., Maniega, S. M., et al. (2015). Rationale, design and methodology of the image analysis protocol for studies of patients with cerebral small vessel disease and mild stroke. *Brain Behav.* 5:e00415. doi: 10.1002/brb3.415

Ioffe, S., and Szegedy, C. (2015). Batch normalization: accelerating deep network training by reducing internal covariate shift. *arXiv* arXiv:1502.03167

Jagust, W. J., Zheng, L., Harvey, D. J., Mack, W. J., Vinters, H. V., Weiner, M. W., et al. (2008). Neuropathological basis of magnetic resonance images in aging and dementia. *Ann. Neurol.* 63, 72–80. doi: 10.1002/ana.21296

Karargyros, A., and Syeda-Mahmood, T. (2018). "Saliency U-Net: a regional saliency map-driven hybrid deep learning network for anomaly segmentation," in *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* (Houston, TX), 105751T.

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization. *arXiv* arXiv:1412.6980.

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 3431–3440.

Lopez, M. M., and Ventura, J. (2017). "Dilated convolutions for brain tumor segmentation in mri scans," in *International MICCAI Brainlesion Workshop* (Quebec City, QC: Springer), 253–262.

Lutkenhoff, E. S., Rosenberg, M., Chiang, J., Zhang, K., Pickard, J. D., Owen, A. M., et al. (2014). Optimized brain extraction for pathological brains (optibet). *PLoS ONE* 9:e115551. doi: 10.1371/journal.pone.0115551

Moeskops, P., Veta, M., Lafarge, M. W., Eppenhof, K. A., and Pluim, J. P. (2017). "Adversarial training and dilated convolutions for brain mri segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Quebec City, QC: Springer), 56–64.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., et al. (2005). Ways toward an early diagnosis in alzheimer's disease: the alzheimer's disease neuroimaging initiative (adni). *Alzheimer's Dement.* 1, 55–66. doi: 10.1016/j.jalz.2005.06.003

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Haifa), 807–814.

Najibi, M., Singh, B., and Davis, L. S. (2018). Autofocus: Efficient multi-scale inference. *arXiv* arXiv:1812.01600.

Quellec, G., Lamard, M., Cozic, M., Coatrieux, G., and Cazuguel, G. (2016). Multiple-instance learning for anomaly detection in digital mammography. *IEEE Trans. Med. Imaging* 35, 1604–1614. doi: 10.1109/TMI.2016.2521442

Rachmadi, M. F., del C. Valdés Hernández, M., and Komura, T. (2018a). "Transfer learning for task adaptation of brain lesion assessment and prediction of brain abnormalities progression/regression using irregularity age map in brain MRI," in *PRedictive Intelligence in MEdicine - First International Workshop, PRIME 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings*, 85–93.

Rachmadi, M. F., Hernández, M. V., Li, H., Guerrero, R., Meijboom, R., Wiseman, S., et al. (2019). Limited one-time sampling irregularity map (LOTS-IM) for automatic unsupervised assessment of white matter hyperintensities and multiple sclerosis lesions in structural brain magnetic resonance images. *bioRxiv*. 334292. doi: 10.1101/334292

Rachmadi, M. F., Valdés-Hernández, M. d. C., and Komura, T. (2017). "Voxel-based irregularity age map (iam) for brain's white matter hyperintensities in mri," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (Bali: IEEE), 321–326.

Rachmadi, M. F., Valdés-Hernández, M. d. C., and Komura, T. (2018b). "Automatic irregular texture detection in brain mri without human supervision," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018*, eds A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger (Cham: Springer International Publishing), 506–513.

Raz, N., Yang, Y., Dahle, C. L., and Land, S. (2012). Volume of white matter hyperintensities in healthy adults: contribution of age, vascular risk factors, and inflammation-related genetic variants. *Biochim. et Biophys. Acta Mol. Basis Dis.* 1822, 361–369. doi: 10.1016/j.bbadis.2011.08.007

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," *International Conference on Medical Image Computing and Computer-Assisted Intervention 2015: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*, 234–241. doi: 10.1007/978-3-319-24574-4-28

Scheltens, P., Barkhof, F., Valk, J., Algra, P., Hoop, R. G. V. D., Nauta, J., et al. (1992). White matter lesions on magnetic resonance imaging in clinically diagnosed alzheimer's disease: evidence for heterogeneity. *Brain* 115, 735–748.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging* (Boone, NC: Springer), 146–157.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D. L., Collins, D. L., et al. (2011). Evaluating intensity normalization on mris of human brain with multiple sclerosis. *Med. Image Anal.* 15, 267–282. doi: 10.1016/j.media.2010.12.003

Yu, F., and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv* arXiv:1511.07122.

Yu, F., Koltun, V., and Funkhouser, T. (2017). Dilated residual networks. *Comput. Vision Patt. Recogn.* 1:2. doi: 10.1109/CVPR.2017.75

Zhang, C., Luo, W., and Urtasun, R. (2018). "Efficient convolutions for real-time semantic segmentation of 3d point clouds," in *2018 International Conference on 3D Vision (3DV)* (Verona: IEEE), 399–408.

Zhang, Y., Chen, W., Chen, Y., and Tang, X. (2018). A post-processing method to improve the white matter hyperintensity segmentation accuracy for randomly-initialized u-net (Verona: VR). *arXiv*. arXiv:1807.10600.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Neuroinformatics

ORIGINAL RESEARCH
published: 02 July 2019
doi: 10.3389/fninf.2019.00048

Parkinson's Disease Detection Using Isosurfaces-Based Features and Convolutional Neural Networks

Andrés Ortiz[1], *Jorge Munilla*[1]*, *Manuel Martínez-Ibañez*[1], *Juan M. Górriz*[2], *Javier Ramírez*[2] *and Diego Salas-Gonzalez*[2]

[1] *Department of Communications Engineering, Universidad de Málaga, Malaga, Spain,* [2] *Department of Signal Theory, Networking and Communications, University of Granada, Granada, Spain*

Computer aided diagnosis systems based on brain imaging are an important tool to assist in the diagnosis of Parkinson's disease, whose ultimate goal is the detection by automatic recognizing of patterns that characterize the disease. In recent times Convolutional Neural Networks (CNN) have proved to be amazingly useful for that task. The drawback, however, is that 3D brain images contain a huge amount of information that leads to complex CNN architectures. When these architectures become too complex, classification performances often degrades because the limitations of the training algorithm and overfitting. Thus, this paper proposes the use of isosurfaces as a way to reduce such amount of data while keeping the most relevant information. These isosurfaces are then used to implement a classification system which uses two of the most well-known CNN architectures, LeNet and AlexNet, to classify DaTScan images with an average accuracy of 95.1% and AUC = 97%, obtaining comparable (slightly better) values to those obtained for most of the recently proposed systems. It can be concluded therefore that the computation of isosurfaces reduces the complexity of the inputs significantly, resulting in high classification accuracies with reduced computational burden.

Keywords: deep learning, isosurfaces, Parkinson's disease, convolutional neural networks, computer-aided diagnosis

OPEN ACCESS

Edited by:
Jesus M. Cortes, BioCruces Health Research Institute, Spain

Reviewed by:
Li Su, University of Cambridge, United Kingdom
Ruxandra Stoean, University of Craiova, Romania

***Correspondence:**
Jorge Munilla
munilla@ic.uma.es

Received: *26 February 2019*
Accepted: *11 June 2019*
Published: *02 July 2019*

Citation:
Ortiz A, Munilla J, Martínez-Ibañez M, Górriz JM, Ramírez J and Salas-Gonzalez D (2019) Parkinson's Disease Detection Using Isosurfaces-Based Features and Convolutional Neural Networks. Front. Neuroinform. 13:48. doi: 10.3389/fninf.2019.00048

1. INTRODUCTION

Parkinson's Disease (PD) is a progressive and chronic neurodegenerative disorder of the central nervous system that affects movement. PD increases its occurrence with age and, currently, has a prevalence between 1 and 3% in the population over 65 years of age, becoming the second most common neurodegenerative disorder after the Alzheimer's disease. The origin of the disease has been not determined yet but it is related to the loss of dopaminergic neurons, which causes reduced quantities of dopamine transporters in the striatum (Simuni and Rajesh, 2009). In fact, dopaminergic neurons produce dopamine, a neurotransmitter, in the substantia nigra and and it is transported to the striatum, composed by caudate and putamen, through the nigrostriatal pathway.

To date there is no cure for PD but early diagnosis allows limiting the rate of progression by applying effective management and may help to develop new therapeutic methods. Diagnosis of PD is usually based on clinical examinations that analyze different motor symptoms such as tremor, bradykinesia, rigidity and postural instability (Eckert et al., 2007), along with the response

to levadopa. Levadopa is a chemical product that converts to dopamine so that PD is confirmed whether symptoms reduce after levadopa is administered during a period of time. However, PD can be confused with other parkinsonian syndromes and in the early stages of the disease symptoms are still mild and the response to levadopa are not so clear, which may result in difficult diagnosis. As a consequence, functional neuroimaging are then usually used to improve the early diagnosis of the disease.

Single Photon Emission Tomography (SPECT) using the $^{123}I - ioflupane$ radiotracer (also known by its tradename DaTSCAN) is commonly used for diagnosis of PD. DaTSCAN binds to the dopaminergic transporters at the striatum, allowing to measure quantitatively the amount of DaTSCAN in this region. DaT SPECT or DaTSCAN imaging results in multiple grayscale images captured by a gamma camera rotated through 360° around the body where the intensity of each pixel is directly correlated with the presence of radiotracers registered by the gamma camera. These 2D projections are then reconstructed to produce a 3D image. Comparing to healthy individuals, the resulting image for PD patients displays lower intensity and/or asymmetry in the striatal region. This way, DaTSCAN can be used to differentially diagnose PD with respect to normal or other diseases presenting similar symptons (NC) by detecting dopaminergic deficits.

In recent years, different works have analyzed DaTSCAN images for use in the clinic as an aid to visual reporting. Thus, a range of of semi-quantification methods can be found in the literature (Taylor and Fenner, 2017). These methods compute SBRs (Striatal Binding Ratios) from both, with and without consideration of the caudates, using different methods and establishing certain limits and likelihood of disease being present. The clinician must eventually interpret the results to come to an overall decision. At this point, machine learning algorithms can be used to help with such decision. Machine learning algorithms can combine multiple input variables describing different features to produce a single value that helps the clinician. These methods search statistical differences between two groups, PD and control (NC), using statistical learning (Rojas et al., 2013; Martínez-Murcia et al., 2014a; Martinez-Murcia et al., 2016b; Khedher et al., 2015; Pereira et al., 2015; Badoud et al., 2016). Although there are others, such as Naïve-Bayes (Towey et al., 2011) or logistic lasso (Tagare et al., 2017), in line with general trends, SVM, with linear or radial basis function kernel, has been the most commonly employed tool, and in the last years the use of methods based on artificial neural networks (ANN) have gained popularity.

The development of novel architectures and effective training algorithms has enabled to use multi-layer neural networks or deep neural networks (aka deep learning) for a wide range of applications (LeCun et al., 2015), such as speech recognition (Hinton et al., 2012), drug discovery (Chen et al., 2018) and genomics (Alipanahi et al., 2015), but it is in the field of computer vision and image classification where deep learning, and particularly convolutional neural networks (CNN), has undergone a real revolution of the state of the art (LeCun et al., 2015). CNNs are biologically-inspired models that resemble the human vision system, computing image features at different abstraction levels by means of the convolution operator, which is subsequently applied to the response of the previous layer (Rawat and Wang, 2017). Nowadays, these architectures have practically reached, or even surpassed, human-level performance in object recognition (Kheradpisheh et al., 2016). Two of the most famous CNN architectures are LeNet-5 (LeCun et al., 1998) and AlexNet (Krizhevsky et al., 2012). They have been well-studied and provide good results compared to other machine learning algorithms and even more complex CNNs. In fact, deeper networks (e.g., Inception Szegedy et al., 2015), with higher number of abstraction levels, allow computing more complex features, but they also result much more complex to train. This causes that the performances degrade because the limitations of the training algorithms (He et al., 2016) and that the architectures tend to be overffited. Thus, although deeper architectures have the potential to outperform simpler LeNet-5 and AlexNet, this cannot be always achieved and even so, the gain in accuracy may imply a considerable higher computational burden that may not be always justified (Martinez-Murcia et al., 2018).

This work analyzes DaTSCAN (3D) images and identifies features which are suitable for being used in a computer-aided classification system intended to classify between positive and negative cases of PD. In particular, this is realized through the identification of isosurfaces and the extraction of descriptive features from these by using CNN architectures based on LeNet-5 and AlexNet. Isosurfaces connect voxels that have the specified intensity or value, much the way contour lines connect points of equal elevation. This work culminates in the implementation of a classification system which uses supervised learning through CNN architectures to classify DaTSCAN images with an average accuracy of 95.1%. Sensitivity and specificity of the system have also been calculated resulting at an average of 95.5% and 94.8%, respectively.

After this introduction, the rest of the paper is structured as follows. Section 2 reviews related works for PD diagnosis. Section 3 shows details on the database used in this work, extracted from the Parkinson Progression Neuroimaging Initiative (PPMI, RRID:SCR_006431) database, and the applied preprocessing. Then, section 4 describes the computing of isosurfaces, the analyzed architectures and their training process. Section 5 presents and discusses the classification results using data from the PPMI. And finally, section 6 shows the conclusions drawn from this work along with its practical applicability.

2. RELATED WORK

The high spatial and color resolution provided by current neuroimaging systems has prompted them to become the main diagnosis tool for neurodegenerative disorders. Thus, DaTSCAN SPECT imaging is used routinely for the diagnosis of PD through the evaluation of deficits of dopamine transporters of the nigrostriatal pathway. However, the visual assessment of these images to come to a final diagnostic is, even for expert clinician, a time-consuming and complicate task, which requires having into account many variables. Machine learning algorithms, which allow combining different types of inputs to produce a result, can

potentially overcome this problem. Additionally, the vast amount of information contained in DatSCAN images requires the use of computer aided tools to be exploited, allowing to find complex, disease-related patterns to increase the diagnosis accuracy. We review next the main computer-based techniques proposed in this framework.

Two of the first works to analyze the possibilities of machine learning algorithms with DaTSCAN were Palumbo et al. (2010) and Towey et al. (2011). The former compared a probabilistic neural network (PNN) with a classification tree (CIT) to differentiate between PD and essential tremor. Striatal binding ratios for caudate and putamina on 3 slices were used as image features. The latter used Naïve-Bayes with PCA decomposition of the voxels in the striatal region. These were followed for a series of works where SVMs were used as the main classifier tool, with linear or RBF kernel and different image features. Illán et al. (2012) and later Oliveira and Castelo-Branco (2015) used voxel-as-features; i.e., image voxel intensities are used directly as features. Segovia et al. (2012) used a Partial Least Square (PLS) scheme to decompose DaT images into scores and loading. Then, the scores with the highest Fisher Discriminant Ratios were used as feature for the SVM. Khedher et al. (2015) also used PLS. Rojas et al. (2013) proposed the use of 2D empical mode decomposition to split DaTSCAN images into different intrinsic mode functions, accounting for different frequency subbands. The components were used to select features related to PD that clearly differentiate them from NC, allowing an easy visual inspection. Martínez-Murcia et al. (2014a) decomposed the DaTSCAN images into statistically independent components which revealed patterns associated to PD. Moreover, in this approach, image voxels were ranked by means of their statistical significance in class discrimination. A more recent approach also based on multivariate decomposition techniques is proposed in Ortiz et al. (2018), where the use of functional principal component analysis on 3D images is proposed. This is addressed by sampling the 3D images using fractal curves in order to transform the 3D DatSCAN images into 1D signals, preserving the neighborhood relationship among voxels. Striatal binding ratios for both caudates and putamina were used in Prashanth et al. (2014), Palumbo et al. (2014), and Bhalchandra et al. (2015). Martínez-Murcia et al. (2014b) proposed the extraction of 3D textural-based features (Haralick texture features) for the characterization of the dopamine transporters concentration in the image. And finishing with those based on SVM, Badoud et al. (2016) used univariate (voxel-wise) statistical parametric mapping and multivariate pattern recognition using linear discriminant classifiers to differentiate among different Parkinsonian syndromes.

More recently, methods based on neural networks, especially deep learning-based methods, have paved the way to discover complex patterns and, consequently, to outperform the diagnosis accuracy obtained by classical statistical methodologies (Ortiz et al., 2016; Martinez-Murcia et al., 2017). The use of models containing stacks of layers composed of a large number of units that individually perform simple operations allows to compute models containing a large number of parameters. Moreover, these massively parallelized architectures are able to discover very complex patterns in the data by a learning process formulated as an optimization problem. Zhang and Kagen (2017) proposes a classifier based on a single layer neural network and voxel-as-features from different slices. Martinez-Murcia et al. (2017) and Martinez-Murcia et al. (2018) propose the use of Convolutional Neural Networks (CNN) to discover patterns associated to PD. Increasing the accuracy requires the use of deeper networks, but this increment also makes the network prone to overfitting and push the training algorithms to their performance limits. Thus, architectures combining more elaborated blocks such as in He et al. (2016) have been also proposed to effectively increase the number of layers.

In this work, we describe a classifier based on the well-known CNNs LeNet-5 and AlexNet where the image features used to train them are isosurfaces computed from the regions of interest. The computation of isosurfaces reduces the complexity of the inputs significantly which results in high classification accuracies with reduced computational burden.

3. MATERIALS

3.1. Database

Data used in the preparation of this article was obtained from the PPMI (Parkinson's Progression Markers Initiative, RRID:SCR_006431). PPMI is an observational clinical study to verify progression markers in PD. For up-to-date information on the study, visit https://www.ppmi-info.org/. The images in this database were imaged $4 + 0.5$ h after the injection of between 111 and 185 MBq of DaTSCAN. Raw projection data are acquired into a 128×128 matrix stepping each 3 degrees for a total of 120 projection into two 20% symmetric photopeak windows centered on 159 KeV and 122 KeV with a total scan duration of approximately 30–45 min (The Parkinson Progression Markers Initiative, 2010).

A total of $N = 269$ DaTSCAN images from this database were used in the preparation of the article. Specifically, the baseline acquisition from 158 subjects suffering from PD and 111 normal controls (NC) was used.

3.2. Spatial Normalization

Spatial normalization is frequently used in neuroimaging studies. It eliminates differences in shape and size of brain, as well as local inhomogeneities due to individual anatomic particularities. It is particularly key in group analysis, where voxel-wise differences are analyzed and quantified (Martinez-Murcia et al., 2016a). In this procedure, individual images are mapped from their individual subject space (image space) to a common reference space, usually stated using a template. The mapping involves the minimization of a cost function that quantifies the differences between the individual image space and the template. The most frequent template is the Montreal Neurological Institute (MNI), set by the International Consortium for Brain Mapping (ICBM) as its standard template, currently in its version ICBM152 (Mazziotta et al., 2001), an average of 152 normal MRI scans in a common space using a nine-parameter linear transformation. A particular

case of affine transformation is the similarity transformation, where only scale, translation and rotation are applied. This is often used for motion correction and reorientation of brain images with respect to a reference, and is frequently performed automatically on many imaging equipment. The DaTSCAN images from the PPMI dataset are roughly realigned. We will refer to this as non-normalized (given that it is only a similarity transformation that preserves shape) or "original." We further preprocessed the images using the SPM12 (Functional Imaging Laboratory of the University College London, 2012) New Normalize procedure with default parameters, which applies affine and local deformations to achieve the best warping of the images and a custom DaTSCAN template defined in Salas-Gonzalez et al. (2015).

Finally, the regions of interests, those which reveal dopaminergic activity, were selected. As a result, the images of original size of (95, 69, 79) were converted into images of size (29, 25, 41). This means passing from 498,800 to 29,795 voxels, a diminution of 94%, which reduces dramatically the complexity of the system without losing almost relevant information since beyond the elected area the intensity values of most of the pixels for both groups is very low or zero.

3.3. Intensity Normalization

Intensity normalization is an important step to ensure that the same intensity levels corresponds to similar drug uptakes, so that intensities can be compared as an indirect measure of the neurophysical activity. Similar intensity values should indicate similar drug uptakes and, as a consequence, differences in these values may reveal different pathologies (Martínez-Murcia et al., 2012; Segovia et al., 2012; Padilla et al., 2015).

This paper uses Integral Normalization (Illán et al., 2012):

$$\hat{\mathbf{I}}_i = \mathbf{I}_i / I_{n,i}, \tag{1}$$

where $\mathbf{I}_i$ is the image of the *ith* subject in the dataset, $\hat{\mathbf{I}}_i$ is the normalized image, and I_n is an intensity normalization value that is computed independently for each subject as the mean of the whole image (in an approximation of the integral). Sometimes, for Parkinson studies, I_n is set to the average of the brain without the striatum; although the influence of this is small and it can be approximated by the mean of the whole image. Finally, in this work, the resulting values are further normalized between 0 and 1.

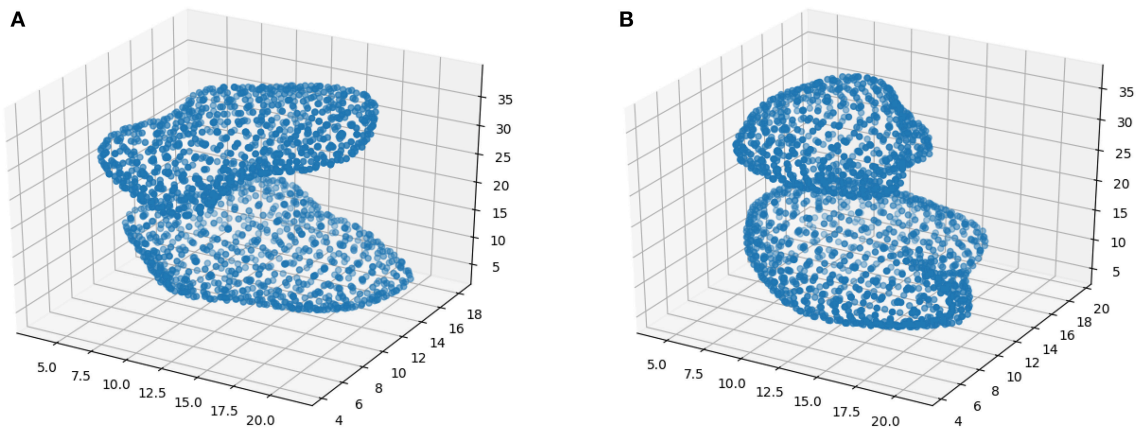

FIGURE 1 | Examples of isosurfaces with threshold = 0.5 for a NC subject **(A)** and PD patient **(B)**.

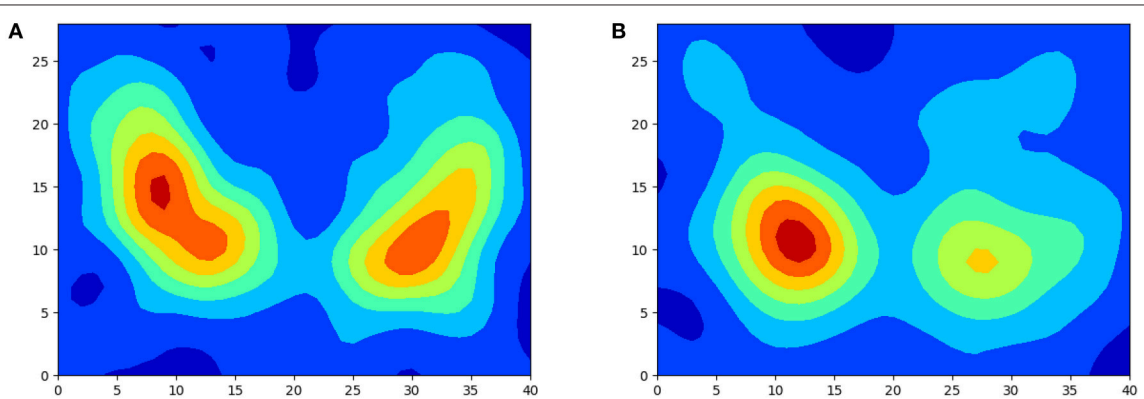

FIGURE 2 | Examples of isolines with different threshold for a NC subject **(A)** and PD patient **(B)**.

4. METHODS

4.1. Feature Extraction Using Isosurfaces

DaTSCAN SPECT images contains an enormous amount of information. The approach of using voxels-as-features has been adopted by different works (Illán et al., 2012; Badoud et al., 2016) reporting modest classification results around 70–75% and suggesting that better results can be achieved by using more refined techniques that focus on the significant information that lies in such images. When CNN are used, as mentioned in the previous sections, this can be explained by a reduction in the complexity that results in lower computational burden (Martinez-Murcia et al., 2018), more efficient training algorithms (He et al., 2016) and less proneness to overfitting. The extraction and selection of features is therefore one of the most determinant processes, and maybe the most characteristic part, in the definition of a classification method.

For feature extraction, this paper proposes the use of isosurfaces. Isosurfaces connect voxels that have the specified intensity or value much the way contour lines connect points of equal elevation. Roughly, this implies to set a threshold at a certain level and take the surface that envelops the remaining voxels above that threshold. In this work, however, a refined version for computing isosurfaces is used where interpolation is employed instead of just thresholding.

Figure 1 shows two examples of isosurfaces computed with a threshold of 0.5 (intensity is normalized to 1) for a NC subject and a PD patient. Unfortunately, it is difficult to observe in a figure different isosurfaces computed for different thresholds since that with the highest threshold will envelop the rest. As an alternative, when different thresholds are used, isolines are preferred. Isolines are simply 2D slices of the corresponding isosurfaces. In **Figure 2** isolines with different thesholds for a NC subject and a PD patient are represented. The following characteristics can be observed in isosurfaces/isolines: (i) they define closed volumes/areas, (ii) they do not cross each other, (iii) the same threshold can result in several isosurfaces/isolines, and (iv) the proximity between isosurfaces/isolines provides information about intensity gradients; the closer they are, the faster the changes. Regarding the diagnosis of PD, it can be observed in previous figures that isosurfaces and isolines from PD patients, in contrast with those from NC subjects, are characterized by a loss of symmetry between hemispheres.

Feature selection is usually based on either statistical analysis or optimization of the classifier. In the former, previously computed thresholds based on statistical relevance or correlations are set and features are discarded if they are not above such thresholds without considering the performance of the classifier. In the latter, however, features are selected, or discarded, if they improve, or not, the performance of the

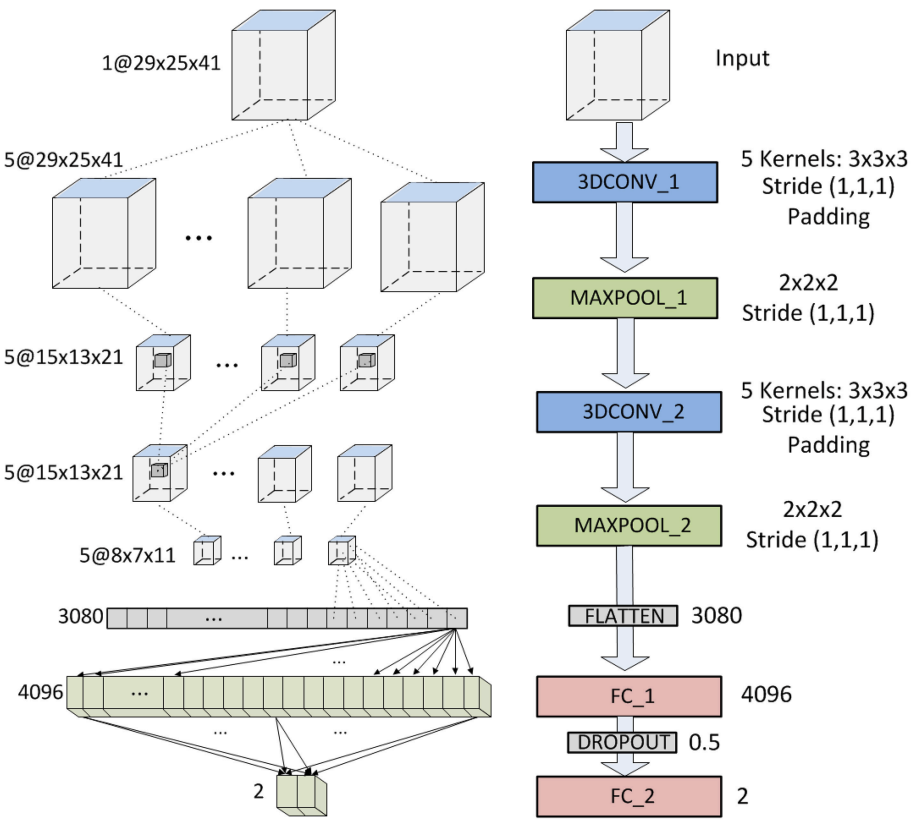

FIGURE 3 | CNN architecture based on LeNet.

classifier. This paper uses this second approach. Classification results using isosurfaces computed with different thresholds have been compared, choosing those that provide the best classification results. More specifically, isosurfaces for thresholds 0.4, 0.5, 0.6, 0.7, and 0.8 have been computed. Then, the two possible options have been analyzed: forward selection, where just one isosurface for a threshold is initially used with the classifier and then others are gradually included if they improve the results; and backward selection, where the whole set of isosurfaces is initially employed to classify and then some of these are removed if their absence does not affect negatively the classification performance.

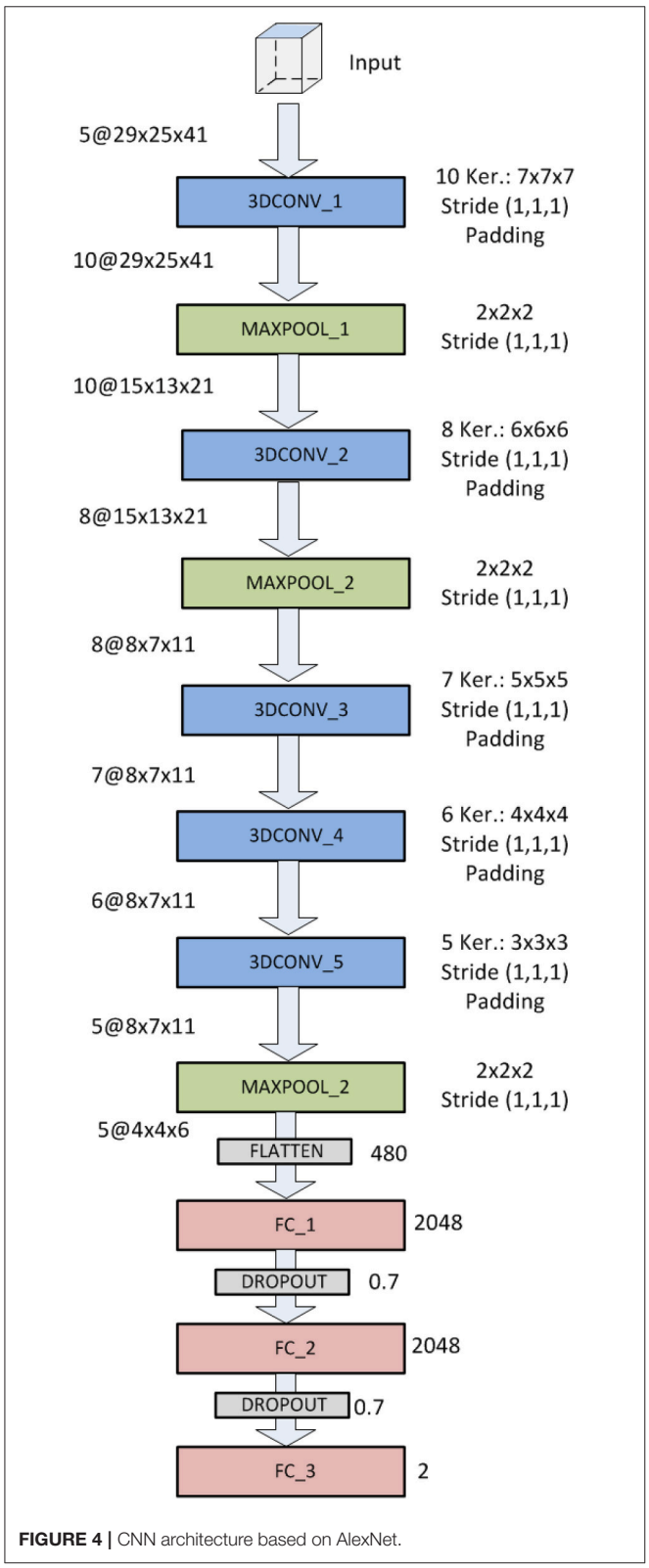

FIGURE 4 | CNN architecture based on AlexNet.

4.2. CNNs for Classification

Method based on neural networks are becoming more and more popular for the development of new early diagnosis tools (Ortiz et al., 2016). More specifically, CNNs have been proposed for the detection of patterns in medical images associated to PD (Martinez-Murcia et al., 2017, 2018). The election and configuration of the CNN architecture are, however, not trivial tasks. In fact, although deeper structures, with higher numbers of layer and units, are potentially more capable of revealing hidden patterns, they are not always advisable because the complexity that they introduce. When a big amount of parameters need to be adjusted, it may result in training problems, overfitting and high computational loads. Thus, apart from preprocessing input data to remove non-significant information and feed CNNs with relevant inputs, the best performances are obtained with balanced architectures; that is, architectures complex enough to reveal the relevant patterns but not so complex that it cannot be conveniently trained with certain guarantees of non-overfitting. In this paper, two 3D versions based on well-known architectures have been tested. The first based on LeNet (LeCun et al., 1998), and then another based on the most powerful AlexNet (Krizhevsky et al., 2012), both of them fed with pre-processed data resulting from the computation of the isosurfaces.

TABLE 1 | Characteristics of the AlexNet-based CNN used.

Layer	Kernel/Window	Output shape	Trainable parameters
Input		29 × 25 × 41	
3D-Conv_1	10@7 × 7 × 7	10@29 × 25 × 41	3,440*
Max_pool_1	2 × 2 × 2	10@15 × 13 × 21	0
3D-Conv_2	8@6 × 6 × 6	8@15 × 13 × 21	17,288
Max_pool_2	2 × 2 × 2	8@8 × 7 × 11	0
3D-Conv_3	7@5 × 5 × 5	7@8 × 7 × 11	7,007
3D-Conv_4	6@4 × 4 × 4	6@8 × 7 × 11	2,694
3D-Conv_5	5@3 × 3 × 3	5@8 × 7 × 11	815
Max_pool_3	2 × 2 × 2	5@4 × 4 × 6	0
Flatten		480	0
FC_1		2,048	985,088
FC_2		2,048	4,196,352
FC_3		2	4,098
		Total	5,216,782

**Computed by a single input volume.*

Our first architecture comprises 7 layers, not counting the input (see **Figure 3**) : 2 convolutional layers (first and third), 2 subsampling layers (second and fourth), 1 flatten layer (fith) and 2 full connected layers (sixth and seventh). The 2 convolutional layers use five 3D-kernels of [3 × 3 × 3] to sweep over the input topologies and transform them into feature maps. Stride

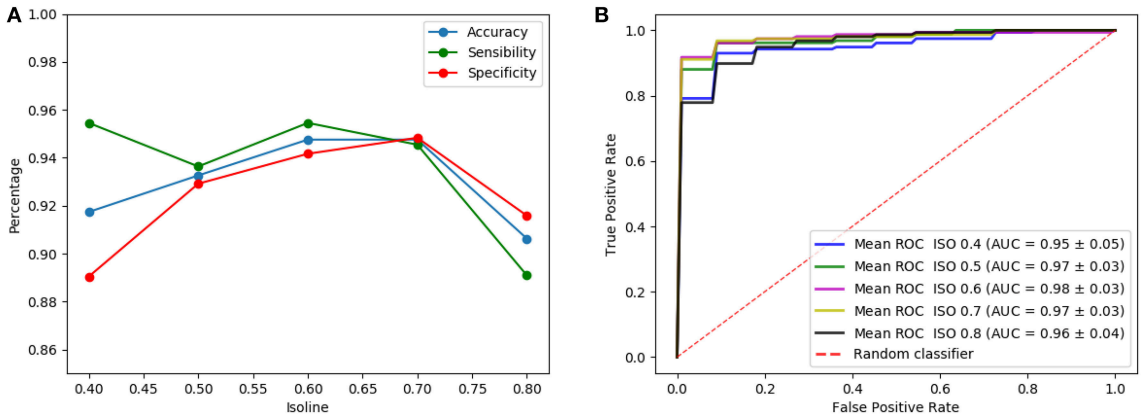

FIGURE 5 | Results of the LeNet-based architecture using as input a single isosurface: sensibilities, sensitivities and accuracies **(A)** and ROC curves **(B)**.

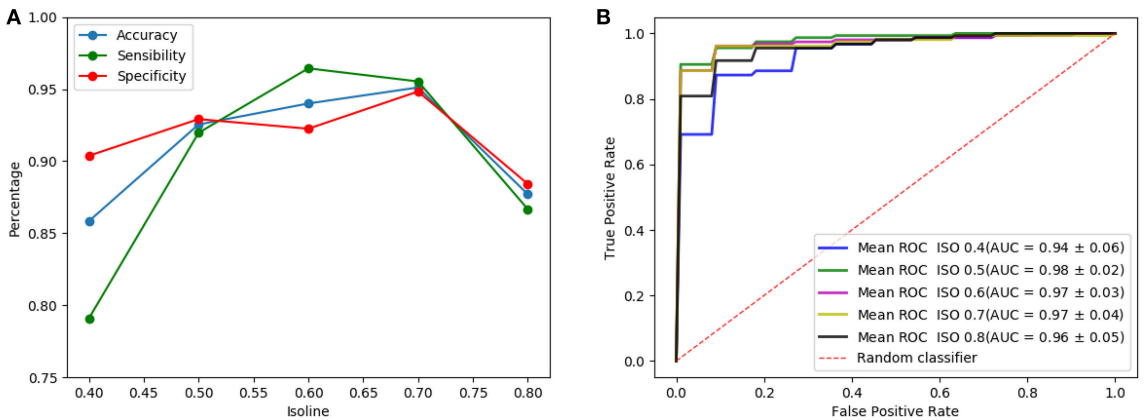

FIGURE 6 | Results of the AlexNet-based architecture using as input a single isosurface: sensibilities, sensitivities and accuracies **(A)** and ROC curves **(B)**.

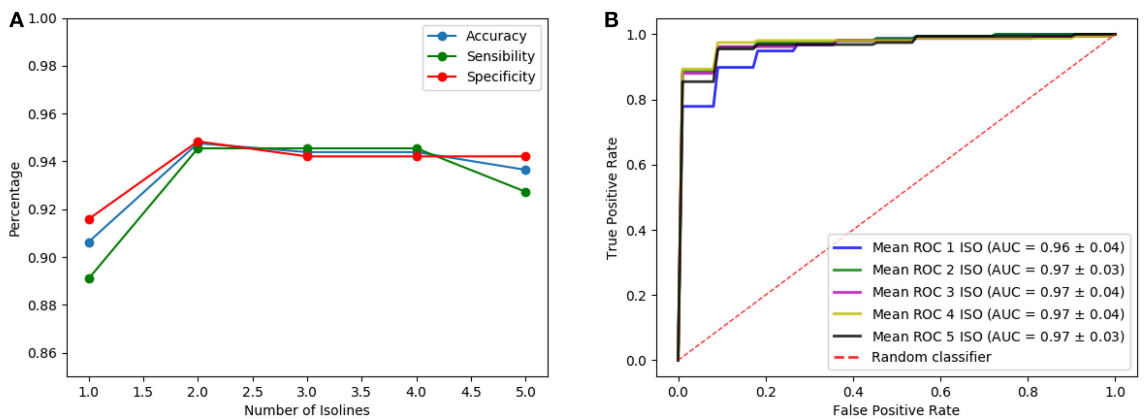

FIGURE 7 | Results of the LeNet-based architecture using as input several isosurfaces: sensibilities, sensitivities and accuracies **(A)** and ROC curves **(B)**.

of (1,1,1) and padding are employed with the convolution so that the output feature maps keep the size of the input. For the second convolutional layer (3DCONV_2), each unit in each feature map is connected to $[3 \times 3 \times 3]$ neighborhoods at identical locations in the entire set of input feature maps. Thus, the number of trainable parameters of these two layers are $3^3 * 5 + 5 = 140$ and $3^3 * 5 * 5 + 5 = 680$, respectively. Note however, that the number of trainable parameters of the first layer increases if several images are introduces simultaneously ($\#params = 3^3 * 5 * \#inputs + 5$). The two subsampling layers apply max-pooling, connecting each unit in the output feature map to $[2 \times 2 \times 2]$ neighborhood in the input feature map. The output is the maximum within the $[2 \times 2 \times 2]$ window. Consequently, the output feature maps have half the number of units in the three dimensions. Sub-sampling reduces the complexity of the CNN and provides invariance to local translations. Once the feature learning phase is completed, using the convolutional and sub-sampling layers, feature maps are flattened into a feature vector. This vector consists of $8 * 7 * 11 * 5 = 3,080$ neurons, and is followed by two fully-connected layers of 4,096 and 2 neurons, respectively. The number of trainable parameters of the last two layers are $3,080 * 4,096 + 4,096 = 12,6190,776$ and $4,096 * 2 + 2 = 8,194$, respectively. Between these two layers there is a dropout interphase with 0.5 dropout probability. The last layer yields the prediction probability using softmax activation. The total number of trainable parameters of this CNN is 12,628,790.

The AlexNet based architecture is shown in **Figure 4**. It comprises 12 layers: 5 (first, third, fourth, fifth and sixth) 3D-convolutional layers, 3 (second, fourth and eight) max-pooling (subsampling) layers, 1 flatten layer (ninth) and 3 fully-connected layers (tenth, eleventh and twelfth). The convolutional layers use 10, 8, 7, 6 and 5 kernels of sizes $[7 \times 7 \times 7]$, $[6 \times 6 \times 6]$, $[5 \times 5 \times 5]$, $[4 \times 4 \times 4]$ and $[3 \times 3 \times 3]$, respectively. Convolutional layers use padding and stride (1,1,1), and output feature maps are connected to every input feature map (not just a subset). The flatten layer has 480 neurons and the three last fully connected layers 2,048, 2,048, and 2, respectively. Between these three fully connected layers there are two dropout interphases with dropout probability of 0.7. The last two-neuron layers uses softmax activation to predict a classification. These characteristics and information about the number of trainable parameters of this CNN are summarized in **Table 1**.

4.3. Evaluation

Classification performance is evaluated by means of the accuracy, sensitivity and specificity. Resulting from these values, Receiver Operating Curves (ROC) and the Areas Under the ROC Curves are also computed. ROC curves comprise the sensitivity and specificity to provide compromise values between these two values, while AUC provides a metric regarding the performance of the classifier.

Classification experiments conducted in this work have been assessed by nested cross-validation (Stone, 1974), with inner and outer loops implementing stratified k-fold cross-validation (k=10) to ensure that the proportion of both classes is preserved in each fold. The inner loop is used to select the features and the outer to determine the generalization ability of the proposed method (Lozano et al., 2017). However, in order to provide a sweep of the performances obtained for different thresholds and to carry out a fair comparison with the optimal one, the results of the outer loop are provided for the different used values (even when they were not the optimal in the inner loop). Estimation of the generalization error by cross-validation will always result in an overestimate in practice, since the entire training set is not used but just a fraction. This overestimate will depend on the slope of the learning curve of the classifier and reduces when k increases.

Standard error is computed from the standard deviation. Cross-validations performed for $k << N$ (where N is the number of samples) allow to estimate the standard deviation of an experiment $CV(\zeta)$. First, the validation error in the j-th fold is averaged as

$$CV_j(\zeta) = \frac{1}{n_j} e_j(\zeta) = \frac{1}{n_j} \sum_{i \in F_j} (y_i - \widehat{f}_\zeta^{-j}(x_i))^2 \qquad (2)$$

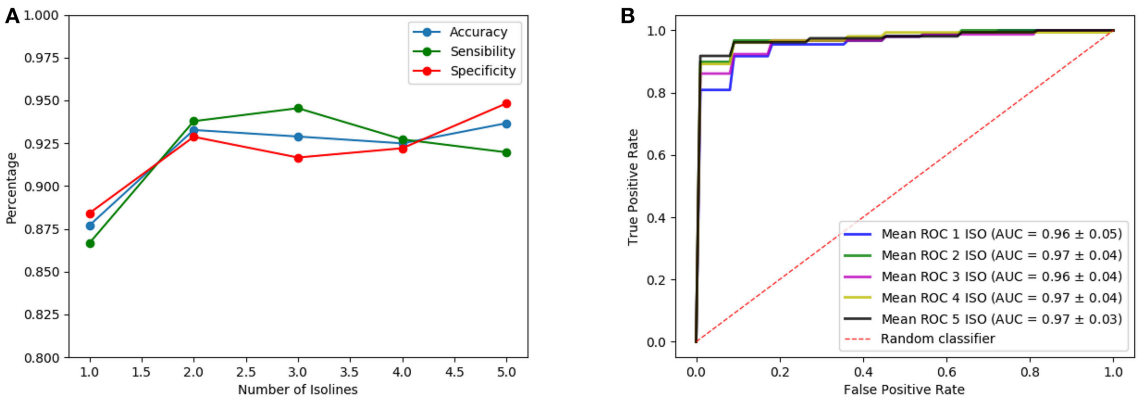

FIGURE 8 | Results of the AlexNet-based architecture using as input several isosurfaces: sensibilities, sensitivities and accuracies **(A)** and ROC curves **(B)**.

where n_j is the number of samples in the j-th fold. Then, the standard deviation of $CV_j(\zeta)$ with $1 \leq j \leq k$ can be computed as:

$$SD(\zeta) = \sqrt{var(CV_1(\zeta), ..., CV_k(\zeta))} \quad (3)$$

where $var(x)$ stands for the variance of the vector x. Finally, the standard error [or standard deviation of $CV(\zeta)$] is computed as:

$$SEM(\zeta) = k^{-\frac{1}{2}} SD(\zeta) \quad (4)$$

5. RESULTS AND DISCUSSION

In this section, we firstly compare classification results when just a single input volume (isosurface) is introduced in the LeNet-based and AlexNet-based architectures. This allows determining which isosurfaces provide more significant information and comparing the performances of both architectures.

Figure 5 shows the results of the LeNet-based architecture for the computed isosurfaces (see section 4.1); **Figure 5A** graphs sensibilities, sensitivities and accuracies, and **Figure 5B** the ROC curves. Likewise, **Figure 6** shows the results for the AlexNet-based architecture. Classification performances increase slightly with the threshold, until this is 0.7. This is explained because the greater the threshold the less the volume captured by the isosurfaces. Thus, as the threshold increases but the chosen volume still contains most of the relevant regions (around the striatum) the performances maintain or improve, since the computational complexity reduces while keeping the significant information. However, for thresholds beyond 0.8, the captured area reduces too much, leaving out relevant regions for the classification and therefore decreasing the performances. As a result, intermediate values of isosurfaces, i.e., 0.5, 0.6, and 0.7, seem to contain the most relevant information providing slightly better classification results for both architectures. On the other hand, there is not a clear difference between the results of the two architectures, both achieving similar performances.

Once the analysis using isosurfaces independently is completed, classification performances obtained when different number of isosurfaces are used as input of the architectures are compared. Note that, although the introduction of more isosurfaces adds more information, it also increases the

TABLE 2 | Classification results using different methods.

Method	Accuracy	Sensitivity	Specificity	AUC
EMD (Rojas et al., 2013)	0.95	0.95	0.94	0.94
Significance M. (Martínez-Murcia et al., 2014a)	0.92	0.95	0.89	0.90
Brahim et al. (2015)	0.92	0.94	0.91	–
VAF	0.8 ± 0.05	0.72 ± 0.17	0.85 ± 0.14	0.87
PCA	0.87 ± 0.04	0.96 ± 0.03	0.86 ± 0.04	0.9
EfPCA (Ortiz et al., 2018)	0.93 ± 0.05	0.97 ± 0.08	0.88 ± 0.05	0.94
LeNet-based	**0.95 ± 0.03**	**0.94 ± 0.04**	**0.95 ± 0.04**	**0.97**
AlexNet-based	**0.95 ± 0.03**	**0.95 ± 0.05**	**0.95 ± 0.04**	**0.97**

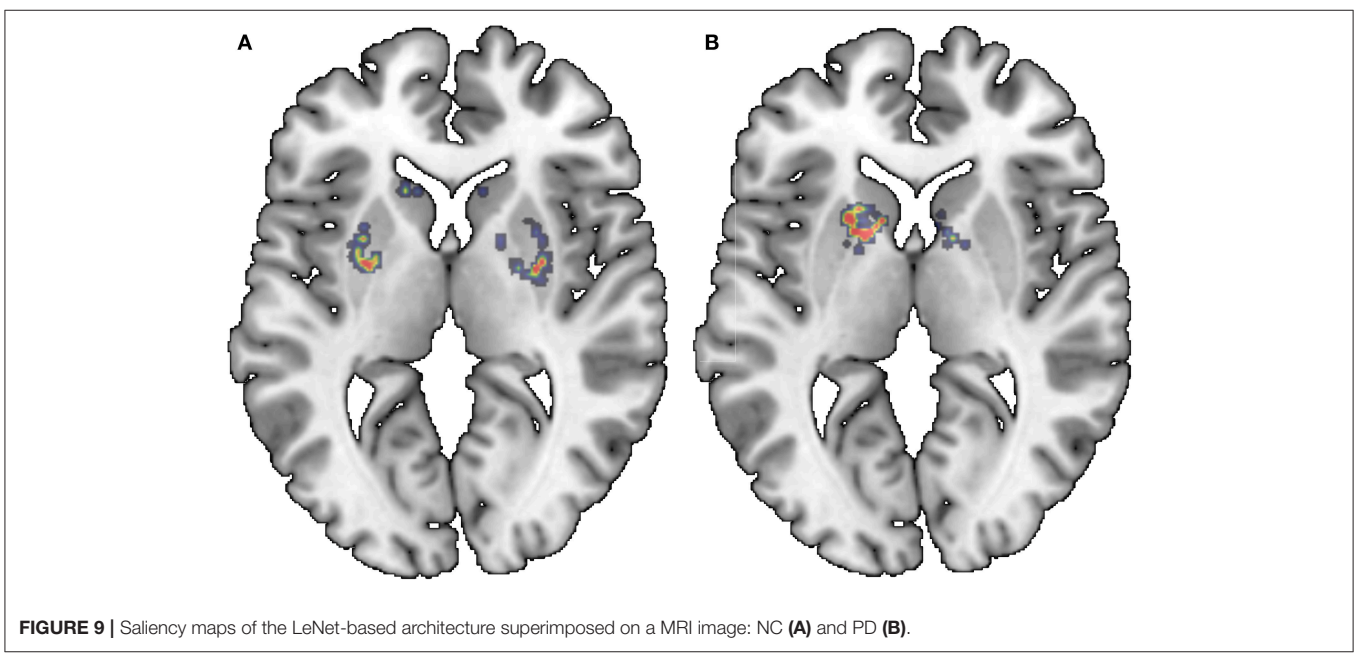

FIGURE 9 | Saliency maps of the LeNet-based architecture superimposed on a MRI image: NC **(A)** and PD **(B)**.

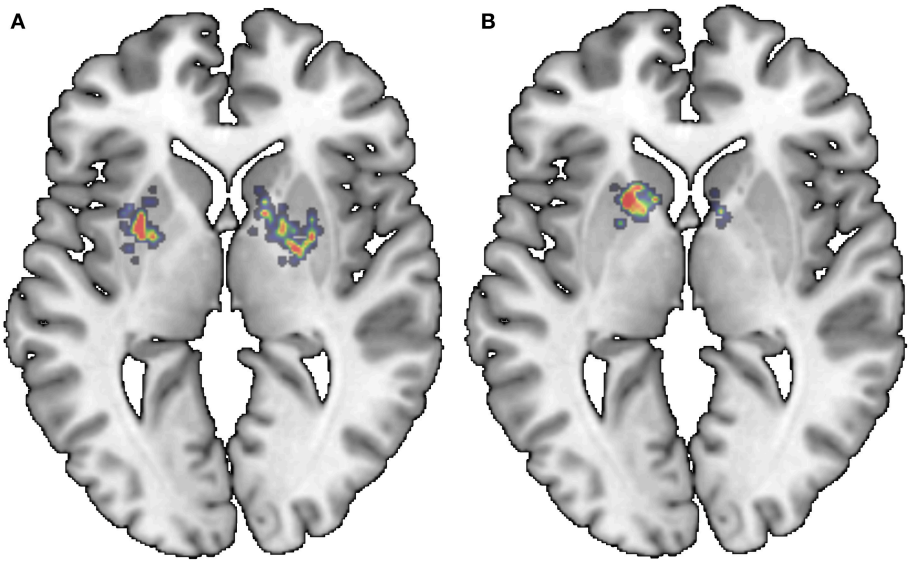

FIGURE 10 | Saliency maps of the AlexNet-based architecture superimposed on a MRI image: NC **(A)** and PD **(B)**.

complexity of the CNN (number of trainable parameters of the first layer) so that the best results are only obtained when the input has an optimum trade-off between the information it provides and the complexity that it introduces. Thus, many possible combinations of isosurfaces have been tested. One of these tests is shown in **Figures 7**, **8** for the LeNet-based and AlexNet-based architecture, respectively. They show the results for the case where isosurfaces are sequentially added from top level (0.8) to bottom level (0.4); that is, first the isosurface with level 0.8 is used by itself (marked as 1 in the figures), then 0.7 is added (2 in the figures), next 0.6 is also added (3 in the figures) and so on: 0.8+0.7+0.6+0.5 (4 in the figures) and all of them (5 in the figures).

The inputs chosen eventually as providing the best classification results while keeping the complexity as low as possible have been the combination of isosurfaces 0.8 and 0.7 for the LeNet-based architecture and the isosurface 0.7 for the AlexNet-based architecture. They both provide accuracy, sensibility and specificity about 0.95 and AUC = 0.97. These classification performances can be considered as very good when compared with other well-known methods such as VAF (Voxels as Features), PCA (Principal Component Analysis) or EfPCA (Empirical functional PCA), outperforming most methods recently published in the bibliography for the detection of Parkinsonism (Rojas et al., 2013; Martínez-Murcia et al., 2014a; Brahim et al., 2015). **Table 2** collects the different performance classifications, including the typical deviation when available. Additionally, in order to statistically confirm the effectiveness of the proposed method, a statistical hypothesis test (Welch test) has been performed in terms of the AUC. As a result, the statistical significance of the use of isosurfaces along with the LeNet-based architecture when compared with the EfPCA (the next best performing method in the comparison) is confirmed with a p-value of 0.04. By contrast, a p-value of 0.18 is computed when both architectures, LeNet and AleNet based, are compared, which allows to infer that while the use of isosurfaces as a feature extraction method outperforms previous approaches, it is not possible to state if one of the two architectures performs better than the other.

Finally, and for the sake of completeness, the saliency maps for the last layer of the Lenet-based and AlexNet-based architecture are provided. **Figures 9**, **10** show a relevant slice of the saliency maps obtained for both architectures superimposed on a MRI image. Saliency maps use the gradient of output category with respect to input image to determine the regions of the input image that have a greater impact on the output class. Thus, for the Alexnet-based architecture (**Figure 10**), it is observed that for control subjects the most decisive regions are those between the putamen and globus pallidus, while for PD patients, the most important ones are those in the interface between the caudate nucleus and the putamen. Similar regions are found in the case of the LeNet-based classifier. However, in this latter case, for the control subjects, sparser regions are marked in the figure (**Figure 9**), while for PD patients, it again shows as the most determinant regions the interface between the caudate nucleus and the left putamen. These anatomical regions matched with those reported in the literature (Greenberg et al., 2012; Tuite et al., 2013) as linked to the development of the Parkinson's disease, which confirms the use of isosurfaces as an effective means to extract the most relevant information for PD diagnosis.

6. CONCLUSIONS

This paper proposes the use of isosurfaces as a way to extract the relevant information from 3D DatSCAN images so that they can be used as inputs of CNN architectures. As a result, a classification system that uses LeNet-based and AlexNet CNN architectures has been implemented. This

system achieves accuracy of 95.1% and AUC = 97%, providing comparable (slightly better) values to those obtained for recently proposed systems. It can be concluded, therefore, that the computation of isosurfaces reduces the complexity of the inputs significantly while keeping the relevant information, resulting in high classification accuracies with reduced computational burden. Finally, in order to determine which areas of the input brain images has a greater impact on the predicted output class, saliency maps of the last two-neuron layer are also computed.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: "https://www.ppmi-info.org/" (accessed June 18, 2019).

AUTHOR CONTRIBUTIONS

All the authors have been involved in the different phases of the development of this work without being possible to set a clear distinction between the different tasks.

FUNDING

This work was partly supported by the MINECO/FEDER under TEC2015-64718-R, PSI2015-65848-R, PGC2018-098813-B-C32, and RTI2018-098913-B-100 projects. We gratefully acknowledge the support of NVIDIA Corporation with the donation of one of the GPUs used for this research.

ACKNOWLEDGMENTS

Data collection and sharing for this project was funded by the Parkinson's Progression Markers Initiative (PPMI), RRID:SCR_006431. It is a public-private partnership funded by The Michael J. Fox Foundation for Parkinson's and other Research and funding partners, including Abbott, Biogen Idec, F. Hoffman-La Roche Ltd., GE Healthcare, Genentech, and Pfizer Inc.

PPMI is an observational clinical study to verify progression markers in Parkinson's disease. PPMI has emerged as a model for following multiple cohorts of significant interest and is being conducted at a network of clinical sites around the world. The study is designed to establish a comprehensive set of clinical, imaging and biosample data that will be used to define biomarkers of PD progression. Once these biomarkers are defined, they can be used in therapeutic studies, which is the ultimate goal.

PPMI will follow standardized data acquisition protocols to ensure that tests and assessments conducted at multiple sites and across multiple cohorts can be pooled in centralized databases and repositories. The clinical, imaging and biologic data will be easily accessible to researchers in real time through this website. The biological samples collected throughout the course of PPMI will be stored in a central repository that will be accessible to any scientist with promising biomarker leads for the purposes of verifying initial results and assessing correlations to clinical outcomes and other biomarkers. Only with collaborative efforts like PPMI can we efficiently identify and validate biomarker candidates for PD progression.

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M. T., and J Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831-838. doi: 10.1038/nbt.3300

Badoud, S., Ville, D. V. D., Nicastro, N., Garibotto, V., Burkhard, P. R., and Haller, S. (2016). Discriminating among degenerative parkinsonisms using advanced 123i-ioflupane spect analyses. *NeuroImage* 12, 234–240. doi: 10.1016/j.nicl.2016.07.004

Bhalchandra, N. A., Prashanth, R., Roy, S. D., and Noronha, S. (2015). "Early detection of parkinson's disease through shape based features from lt;sup gt;123 lt;/sup gt;i-ioflupane spect imaging," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* (Brooklyn, NY), 963–966.

Brahim, A., Ramírez, J., Górriz, J., Khedher, L., and Salas-Gonzalez, D. (2015). Comparison between different intensity normalization methods in 123I-ioflupane imaging for the automatic detection of parkinsonism. *PLoS ONE* 10:e0130274. doi: 10.1371/journal.pone.0130274

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250. doi: 10.1016/j.drudis.2018.01.039

Eckert, T., Tang, C., and Eidelberg, D. (2007). Assessment of the progression of parkinson's disease: a metabolic network approach. *Lancet Neurol.* 6, 926–932. doi: 10.1016/S1474-4422(07)70245-4

Functional Imaging Laboratory of the University College London (2012). *Statistical Parametric Mapping (SPM12).* Available online at: https://www.fil.ion.ucl.ac.uk/spm/ (accessed June 18, 2019).

Greenberg, D. A., Aminoff, M. J., and Simon, R. P. (2012). *Clinical Neurology*. New York, NY: McGraw-Hill Medical.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778.

Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., rahman Mohamed, A., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* 29, 82–97. doi: 10.1109/MSP.2012.2205597

Illán, I. A., Górriz, J. M., Ramírez, J., Segovia, F., Hoyuela, J. M. J., and Lozano, S. J. O. (2012). Automatic assistance to parkinsons disease diagnosis in datscan spect imaging. *Med. Phys.* 39, 5971–5980. doi: 10.1118/1.4742055

Khedher, L., Ramírez, J., Górriz, J., Brahim, A., and Segovia, F. (2015). Early diagnosis of disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. *Neurocomputing* 151, 139–150. doi: 10.1016/j.neucom.2014.09.072

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Sci. Rep.* 6:32672. doi: 10.1038/srep32672

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, NIPS'12* (Lake Tahoe, NV: Curran Associates Inc.), 1097–1105.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Lozano, F., Ortiz, A., Munilla, J., Peinado, A., and for the Alzheimer's Disease Neuroimaging Initiative (2017). Automatic computation of regions of interest

by robust principal component analysis. application to automatic dementia diagnosis. *Knowl. Based Syst.* 123, 229–237. doi: 10.1016/j.knosys.2017.02.025

Martinez-Murcia, F., Górriz, J., and Ramírez, J. (2016a). *Computer Aided Diagnosis in Neuroimaging, 1st Edn.* Chap. 7. InTech.

Martinez-Murcia, F., Górriz, J., Ramírez, J., Ortiz, A., and for the Alzheimer's Disease Neuroimaging Initiative (2016b). A spherical brain mapping of MR images for the detection of Alzheimer's disease. *Curr. Alzheimer Res.* 13, 575–588. doi: 10.2174/1567205013666160314145158

Martínez-Murcia, F., Górriz, J., Ramírez, J., Puntonet, C., and Salas-González, D. (2012). Computer aided diagnosis tool for Alzheimer's disease based on Mann-Whitney-Wilcoxon U-Test. *Exp. Syst. Appl.* 39, 9676–9685. doi: 10.1016/j.eswa.2012.02.153

Martínez-Murcia, F. J., Górriz, J. M., Ramírez, J., Illán, I. A., and Ortiz, A. (2014a). Automatic detection of parkinsonism using significance measures and component analysis in datscan imaging. *Neurocomputing* 126, 58–70. doi: 10.1016/j.neucom.2013.01.054

Martínez-Murcia, F. J., Górriz, J. M., Ramírez, J., Moreno-Caballero, M., A., and Gómez-Río, M. (2014b). Parametrization of textural patterns in 123i-ioflupane imaging for the automatic detection of parkinsonism. *Med. Phys.* 41:012502. doi: 10.1118/1.4845115

Martinez-Murcia, F. J., Górriz, J. M., Ramírez, J., and Ortiz, A. (2018). Convolutional Neural Networks for neuroimaging in Parkinson's disease: is preprocessing needed? *Int. J. Neural Syst.* 28:1850035. doi: 10.1142/S0129065718500351

Martinez-Murcia, F. J., Ortiz, A., Górriz, J. M., Ramírez, J., Segovia, F., Salas-Gonzalez, D., et al. (2017). "A 3d convolutional neural network approach for the diagnosis of parkinson's disease," in *Natural and Artificial Computation for Biomedicine and Neuroscience*, eds J. M. Ferrández Vicente, J. R. Álvarez-Sánchez, F. de la Paz López, J. Toledo Moreo, and H. Adeli (Cham: Springer International Publishing), 324–333.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., et al. (2001). A probabilistic atlas and reference system for the human brain: international consortium for brain mapping (icbm). *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356, 1293–1322. doi: 10.1098/rstb.2001.0915

Oliveira, F. P., and Castelo-Branco, M. (2015). Computer-aided diagnosis of Parkinson's disease based on [(123)I]FP-CIT SPECT binding potential images, using the voxels-as-features approach and support vector machines. *J. Neural Eng.* 12:26008. doi: 10.1088/1741-2560/12/2/026008

Ortiz, A., Martínez-Murcia, F. J., García-Tarifa, M. J., Lozano, F., Górriz, J. M., and Ramírez, J. (2016). "Automated diagnosis of parkinsonian syndromes by deep sparse filtering-based features," in *Innovation in Medicine and Healthcare*, eds Y. W. Chen, S. Tanaka, R. Howlett, and L. Jain (Puerto de la Cruz: Springer), 249–258.

Ortiz, A., Munilla, J., Martinez-Murcia, F. J., Górriz, J. M., and Ramírez, J. (2018). Empirical functional pca for 3d image feature extraction through fractal sampling. *Int. J. Neural Syst.* 29, 1–22. doi: 10.1142/S0129065718500405

Padilla, P., Górriz, J., Ramírez, J., Salas-González, D., and Illán, I. (2015). Intensity normalization in the analysis of functional datscan spect images: the distribution-based normalization method vs other approaches. *Neurocomputing* 150, 4–15. doi: 10.1016/j.neucom.2014.01.080

Palumbo, B., Fravolini, M. L., Buresta, T., Pompili, F., Forini, N., Nigro, P., et al. (2014). Diagnostic accuracy of Parkinson disease by support vector machine (SVM) analysis of 123I-FP-CIT brain SPECT data: implications of putaminal findings and age. *Medicine* 93:e228. doi: 10.1097/MD.0000000000000228

Palumbo, B., Fravolini, M. L., Nuvoli, S., Spanu, A., Paulus, K. S., Schillaci, O., et al. (2010). Comparison of two neural network classifiers in the differential diagnosis of essential tremor and Parkinson's disease by (123)I-FP-CIT brain SPECT. *Eur. J. Nucl. Med. Mol. Imaging* 37, 2146–2153. doi: 10.1007/s00259-010-1481-6

Pereira, C. R., Pereira, D. R., d. Silva, F. A., Hook, C., Weber, S. A. T., Pereira, L. A. M., et al. (2015). "A step towards the automated diagnosis of parkinson's disease: analyzing handwriting movements," in *2015 IEEE 28th International Symposium on Computer-Based Medical Systems* (Sao Carlos), 171–176.

Prashanth, R., Dutta Roy, S., Mandal, P. K., and Ghosh, S. (2014). Automatic classification and prediction models for early Parkinson's disease diagnosis from spect imaging. *Exp. Syst. Appl.* 41, 3333–3342. doi: 10.1016/j.eswa.2013.11.031

Rawat, W., and Wang, Z. (2017). Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* 29, 2352–2449. doi: 10.1162/neco_a_00990

Rojas, A., Górriz, J., Ramírez, J., Illán, I., Martínez-Murcia, F., Ortiz, A., et al. (2013). Application of empirical mode decomposition (emd) on datscan spect images to explore Parkinson disease. *Exp. Syst. Appl.* 40, 2756–2766. doi: 10.1016/j.eswa.2012.11.017

Salas-Gonzalez, D., Górriz, J. M., Ramírez, J., Illán, I. A., Padilla, P., Martínez-Murcia, F. J., et al. (2015). Building a FP-CIT SPECT brain template using a posterization approach. *Neuroinformatics* 13, 391–402. doi: 10.1007/s12021-015-9262-9

Segovia, F., Górriz, J. M., Ramírez, J., Chaves, R., and Illán, I. Á. (2012). Automatic differentiation between controls and Parkinson's disease DaTSCAN images using a partial least squares scheme and the fisher discriminant ratio. *Front. Art. Intell. Appl.* 243, 2241–2250. doi: 10.3233/978-1-61499-105-2-2241

Simuni, T., and Rajesh, P. (2009). *Parkinson's Diseasy*. New York, NY: Oxford University Press USA.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B (Methodol.).* 36, 111–147.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 1*, (Boston, MA), 1–9.

Tagare, H. D., DeLorenzo, C., Chelikani, S., Saperstein, L., and Fulbright, R. K. (2017). Voxel-based logistic analysis of PPMI control and Parkinson's disease DaTscans. *NeuroImage* 152, 299–311. doi: 10.1016/j.neuroimage.2017.02.067

Taylor, J. C., and Fenner J. W. (2017). Comparison of machine learning and semi-quantification algorithms for (i123)fp-cit classification: the beginning of the end for semi-quantification? *EJNMMI Phys.* 4:29. doi: 10.1186/s40658-017-0196-1

The Parkinson Progression Markers Initiative, P. (2010). *Imaging Technical Operations Manual. 2nd Edn.*

Towey, D. J., Bain, P. G., and Nijran, K. S. (2011). Automatic classification of 123I-FP-CIT (DaTSCAN) SPECT images. *Nucl. Med. Commun.* 32, 699–707. doi: 10.1097/MNM.0b013e328347cd09

Tuite, P. J., Mangia, S., and Michaeli, S. (2013). Magnetic resonance imaging (MRI) in Parkinson's disease. *J. Alzheimer's Dis. Parkinsonism* (Suppl. 1):001. doi: 10.4172/2161-0460.S1-001

Zhang, Y. C., and Kagen, A. C. (2017). Machine learning interface for medical image analysis. *J. Digit. Imaging* 30, 615–621. doi: 10.1007/s10278-016-9910-0

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Aging Neuroscience

ORIGINAL RESEARCH
published: 31 July 2019
doi: 10.3389/fnagi.2019.00194

Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification

Moritz Böhle[1,2‡], *Fabian Eitel*[1,2‡], *Martin Weygandt*[1,3] *and Kerstin Ritter*[1,2*] *on behalf of the Alzheimer's Disease Neuroimaging Initiative*[†]

[1] *Berlin Institute of Health, Charité—Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin, Germany,* [2] *Department of Psychiatry and Psychotherapy, Bernstein Center for Computational Neuroscience, Berlin, Germany,* [3] *Excellence Cluster NeuroCure Berlin, Berlin, Germany*

Deep neural networks have led to state-of-the-art results in many medical imaging tasks including Alzheimer's disease (AD) detection based on structural magnetic resonance imaging (MRI) data. However, the network decisions are often perceived as being highly non-transparent, making it difficult to apply these algorithms in clinical routine. In this study, we propose using layer-wise relevance propagation (LRP) to visualize convolutional neural network decisions for AD based on MRI data. Similarly to other visualization methods, LRP produces a heatmap in the input space indicating the importance/relevance of each voxel contributing to the final classification outcome. In contrast to susceptibility maps produced by guided backpropagation ("Which change in voxels would change the outcome most?"), the LRP method is able to directly highlight positive contributions to the network classification in the input space. In particular, we show that (1) the LRP method is very specific for individuals ("Why does this person have AD?") with high inter-patient variability, (2) there is very little relevance for AD in healthy controls and (3) areas that exhibit a lot of relevance correlate well with what is known from literature. To quantify the latter, we compute size-corrected metrics of the summed relevance per brain area, e.g., relevance density or relevance gain. Although these metrics produce very individual "fingerprints" of relevance patterns for AD patients, a lot of importance is put on areas in the temporal lobe including the hippocampus. After discussing several limitations such as sensitivity toward the underlying model and computation parameters, we conclude that LRP might have a high potential to assist clinicians in explaining neural network decisions for diagnosing AD (and potentially other diseases) based on structural MRI data.

Keywords: Alzheimer's disease, MRI, visualization, explainability, layer-wise relevance propagation, deep learning, convolutional neural networks (CNN)

OPEN ACCESS

Edited by:
James H. Cole,
King's College London,
United Kingdom

Reviewed by:
Taeho Jo,
Indiana University, United States
David Wood,
King's College London,
United Kingdom

***Correspondence:**
Kerstin Ritter
kerstin.ritter@charite.de

[†]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[‡]These authors have contributed equally to this work

Received: *27 February 2019*
Accepted: *15 July 2019*
Published: *31 July 2019*

Citation:
Böhle M, Eitel F, Weygandt M and Ritter K (2019) Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification. Front. Aging Neurosci. 11:194. doi: 10.3389/fnagi.2019.00194

1. INTRODUCTION

In the 2018 World Alzheimer Report, it was estimated that 50 million people worldwide were suffering from dementia and this number was projected to rise to more than 152 million people until 2050. The most common reason for dementia is Alzheimer's disease (AD) accounting for around 60–70 % of dementia cases (WHO, 2017). AD is characterized by abnormal cell death, primarily in the medial temporal lobe. This cell death is thought to be rooted in protein plaques and neurofibrillary tangles, which restrict normal neural function (Bondi et al., 2017). The resulting atrophy is visible in structural magnetic resonance imaging (MRI) data, and derived markers (such as hippocampal volume or gray matter density) have been used to diagnose AD and predict disease progression (Frisoni et al., 2010; Rathore et al., 2017). In the last decade, those markers have frequently been employed in machine learning settings to allow for predictions on an individual level (Klöppel et al., 2008; Orrù et al., 2012; Weiner et al., 2013; Ritter et al., 2015, 2016). However, those expert features usually reflect only one part of disease pathology and the combination with standard machine learning methods, such as support vector machines, do not allow for finding new and potentially unexpected hidden data characteristics that might also be important to describe a disease.

By extracting hierarchical information directly from raw or minimally processed data, deep learning approaches (LeCun et al., 2015) can help to fill a gap here and offer a great potential for improving automatic disease diagnostics. One family of algorithms that perfectly lends itself to perform non-linear feature extraction from image data and their respective classification into disease categories are convolutional neural networks (CNNs), a type of (deep) neural networks optimized for image data. The key idea behind CNNs is inspired by the mechanism of receptive fields in the primate's visual cortex and relates to the application of local convolutional filters for extracting regional information (LeCun and Bengio, 1995). They typically consist of a sequence of convolutional and pooling layers which allow for summarizing key characteristics into feature maps. These feature maps can then be used by a fully-connected layer or any other classifier for solving the primary supervised learning problem (e.g., AD classification). CNNs have been proven to be very successful in a wide range of medical imaging applications (Litjens et al., 2017), including AD detection based on neuroimaging data (e.g., Gupta et al., 2013; Suk et al., 2014; Payan and Montana, 2015; Sarraf and Tofighi, 2016; Korolev et al., 2017; for a review, see Vieira et al., 2017).

Despite this success, automatically learning the features comes at a cost: the decisions of neural networks are notoriously hard to interpret in retrospect. Therefore, deep learning methods, including CNNs, often face the criticism that they are "black-box" (Castelvecchi, 2016). In contrast to some simpler learning algorithms, in particular decision trees, they do not offer a simple and comprehensible explanation; their architecture is complex and consists of several to many layers with hundreds of thousands parameters that need to be trained. In the medical domain, however, it is imperative to base diagnoses and subsequent treatments on an informed decision and not on a single yes/no answer of an algorithm. Therefore, if CNNs should support clinicians in their daily work, ways have to be found to visualize and interpret the network's "decision" (see **Figure 1**). In the last years, a number of suggestions have been made to visualize *what* is actually learned by a CNN. Besides straightforward methods such as the extraction of activations during convolution or the visualization of weights, among the most well-known techniques for visualization are the sensitivity analysis by Simonyan et al. (2013), guided backpropagation by Springenberg et al. (2014), the deep visualization toolbox of Yosinski et al. (2015) based on regularized optimization, and the deconvolution and occlusion method by Zeiler and Fergus (2014). In Alzheimer's research only a very few studies exist that looked into such visualization methods (Esmaeilzadeh et al., 2018; Rieke et al., 2018; Yang et al., 2018).

Most promising for the use in the medical imaging domain is the generation of an individual heatmap for each patient, which lies in the same space as the input image and indicates the importance of each voxel for the final (individual) classification decision. By allowing for a human-guided, intuitive investigation of what drives the classifier to come to a certain classification decision, individual heatmaps hold great potential in assisting and understanding diagnostic decisions performed by deep neural networks. However, for any visualization method that produces heatmaps, it is very important to understand how they are computed and what their limitations are. In natural images, for example, it has been argued that methods relying on gradients (e.g., sensitivity analysis or guided backpropagation) only measure the susceptibility of the output to changes in the input and might not necessarily coincide with those areas on which the network bases its decision. A powerful method to overcome this limitation is layer-wise relevance propagation (LRP, Bach et al., 2015), which decomposes the network's output score (e.g., for AD) into the individual contributions of the input neurons while keeping the total amount of relevance constant across layers (conservation principle). In contrast to showing "susceptibility maps" as gradient-based methods, the heatmap does not rely on gradients, but takes into account model parameters (i.e., weights) and neuron activations (Bach et al., 2015; Samek et al., 2015). By this, the heatmaps are less prone to group effects in the data. Intuitively, LRP has the potential to answer the question "what speaks for AD in this particular patient?" as opposed to "which change in voxels would change the outcome most?" addressed in gradient-based approaches. In terms of explainability, LRP has been shown to be superior to those gradient methods and deconvolution methods in three natural imaging data sets (Samek et al., 2015). In cognitive neuroscience, the LRP method has been recently applied to single-trial EEG and functional MRI classification (Sturm et al., 2016; Thomas et al., 2018). To the best of our knowledge, it has so far not been applied in clinical disease classification based on structural MRI data.

In this study, we use LRP to explain individual classification decisions for AD patients and healthy controls (HCs) based on a CNN trained on structural MRI data (T1-weighted MPRAGE) from the Alzheimer's Disease Neuroimaging Initiative

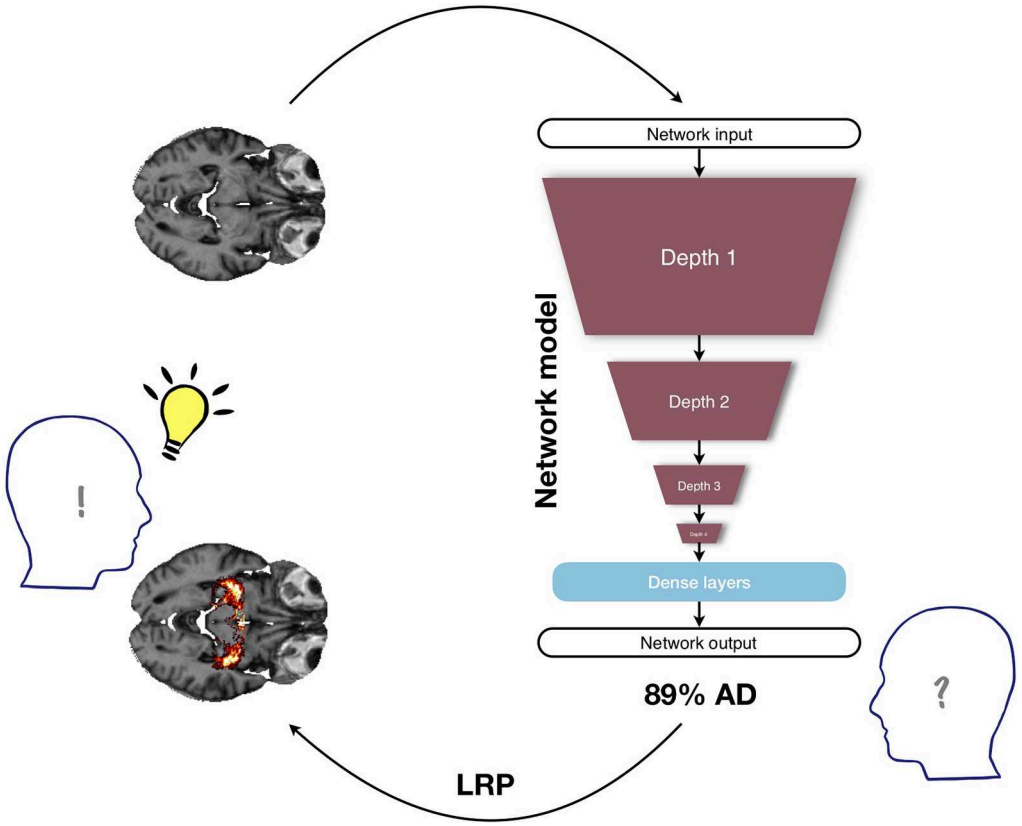

FIGURE 1 | Illustration of the benefit of visualization in a deep learning framework for diagnosing Alzheimer's disease (AD) based on structural MRI data. Deep neural networks are often criticized for being non-transparent, since they usually provide only one single class score as output and do not explain what has led to this particular network decision; in this example, the MRI input is classified as belonging to the group of AD patients with a probability of 89%. When no further information is given, the medical expert is not able to base any medical treatment on this number, since the underlying reasons are unclear. The layer-wise relevance propagation method (LRP) might alleviate this problem by additionally providing a heatmap in which the positive contributions to the class score (89% AD) are highlighted. Here, the class score is supplemented by the additional information that in this particular subject AD relevance has been found in the hippocampus, an area known to be affected in AD. By providing a visual explanation, the LRP framework allows the medical expert to make a much more informed decision.

(ADNI[1]). Based on the trained CNN model, we generated LRP heatmaps for each subject in the test set. Importantly, each heatmap indicates the voxel-wise relevance for the particular classification decision (AD or HC). To spot the most relevant regions for AD classification, we computed average heatmaps across AD patients and HCs, which we then further split into correct and wrong classification decisions (i.e., true positives, false positives, true negatives, false negatives). To analyze the relevance in different brain areas according to the Scalable Brain Atlas by Neuromorphometrics Inc. (Bakker et al., 2015), we suggest size-corrected metrics and compared these metrics between LRP and guided backpropagation. We have chosen guided backpropagation as a baseline method because (1) sensitivity analysis is the most common method for generating heatmaps, (2) it results in more focused heatmaps compared to only using backpropagation (Rieke et al., 2018) and (3) it is better comparable to LRP than occlusion methods with respect to our relevance measures. On an individual level, we analyzed the heatmap patterns of single subjects ("relevance fingerprinting") and correlate them with the hippocampal volume as a key biomarker of AD. We show that the LRP heatmaps succeeded in depicting individual contributions to AD diagnosis and might hold great potential as a diagnostic tool.

2. MATERIALS AND METHODS

2.1. Data and Preprocessing

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI, RRID:SCR_003007) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). For up-to-date information, see www.adni-info.org.

[1]http://adni.loni.usc.edu/

We included structural MRI data of all subjects with Alzheimer's disease (AD) and healthy controls (HCs) listed in the "MRI collection - Standardized 1.5T List - Annual 2 year". The subjects in the data set are labeled as AD if the Clinical Dementia Rating (CDR) score (Morris, 1993) was greater than 0.5. HCs are selected as those subjects with a CDR score of 0. In total, we included 969 individual scans (475 AD, 494 HC) of 193 AD patients and 151 HCs (up to three time points). All scans were acquired with 1.5 T scanners at various sites and had undergone gradient non-linearity, intensity inhomogeneity and phantom-based distortion correction. We downloaded T1-weighted MPRAGE scans and non-linearly registered them to the 1mm resolution 2009c version of the ICBM152 reference brain using Advanced Normalization Tools (ANTs[2]). This has been done to (1) ensure a relative alignment across subjects, (2) allow the convolutional neural network to extract more robust features, and (3) be able to analyze the heatmaps in a common space. For the region-wise analysis of heatmaps, we used the Scalable Brain Atlas by Neuromorphometrics Inc. (Bakker et al., 2015) available in SPM12[3]. A list of all areas included can be found in the SPM12 package.

2.2. Convolutional Neural Network Architecture

Convolutional neural networks (CNNs) are neural networks optimized for array data including images or videos (LeCun et al., 2015). In addition to input and output layer, they consist of several hidden layers including convolutional and pooling layers. In convolutional layers, in contrast to fully-connected layers, the weights and the bias terms are shared between all neurons in a given layer for a given filter. This means that each of the neurons applies the same *filter* or *kernel* to the input, but at a different position, usually with a displacement (often called stride) of 1–3 between neighboring neurons. Since these filters are learned via the backpropagation algorithm, CNNs do not rely on hand-crafted features, but can be applied to minimally processed data (LeCun et al., 2015). CNNs have been very successfully applied to a large number of applications including image and speech recognition (Krizhevsky et al., 2012; Abdel-Hamid et al., 2014; Long et al., 2015) as well as medical imaging and AD classification based on MRI data (Gupta et al., 2013; Suk et al., 2014; Payan and Montana, 2015; Sarraf and Tofighi, 2016; Korolev et al., 2017; Litjens et al., 2017; Vieira et al., 2017).

The model in the present study consists of four convolutional blocks followed by two fully-connected layers. Each block features a convolutional layer with f filters ($f = 8, 16, 32, 64$) and filter sizes of $3 \times 3 \times 3$. Every convolutional layer is followed by batch normalization and max pooling with window sizes wxwxw ($w = 2, 3, 2, 3$). The fully-connected layers contain 128 and 2 units respectively and dropout ($p = 40\%$) is applied before each. The final fully-connected layer, which is activated by a softmax function serves as the network output, providing the class scores for HCs (first unit) and AD (second unit) respectively. As an optimizer Adam (Kingma and Ba, 2015) was used with an initial learning rate of 0.0001 and a weight decay of 0.0001. The data was split into a training data set (163 AD patients, 121 HCs; 797 images in total), a validation set for optimizing the hyperparameters (18 AD patients, 18 HCs; 100 images in total) and a test set (30 AD patients, 30 HCs; 172 images in total). To ensure independence between training and test data, we performed the split of the data on the level of patients instead of images. The data was augmented during training by flipping the images along the sagittal axis ($p = 50\%$) and translated along the sagittal axis between –2 and 2 voxels. When the model did not improve for 8 epochs on the validation set, training was stopped. The training epoch (i.e., model checkpoint) with the best validation accuracy (91.00%) was then applied to the test data, resulting in a classification accuracy of 87.96%.

[2]http://stnava.github.io/ANTs/

[3]http://www.fil.ion.ucl.ac.uk/spm/software/spm12/

2.3. Visualization Methods

2.3.1. Layer-Wise Relevance Propagation (LRP)

In the following, we will introduce the Layer-wise Relevance Propagation (LRP) algorithm by Bach et al. (2015). The core idea underlying the LRP algorithm for attributing relevance to individual input nodes is to trace back contributions to the final output node layer by layer. While several different versions of the LRP algorithm exist, they all share the same principle: the total relevance—e.g., the activation strength of an output node for a certain class—is conserved per layer; each of the nodes in layer l that contributed to the activation of a node j in the subsequent layer $l+1$ gets attributed a certain share of the relevance R^j_{l+1} of that node. Overall, the sum over the relevances of all nodes i contributing to neuron j in layer l must sum to R^j_{l+1}, such that the total relevance per layer is conserved:

$$\sum_i R^{i\to j}_{l,l+1} = R^j_{l+1} \tag{1}$$

There are different ways in which the relevance can be distributed over the input nodes i and different rules for how to distribute the relevances have been proposed. In this paper, we used the β-rule (as described in Binder et al., 2016b):

$$R^{i\to j}_{l,l+1} = \left((1+\beta)\frac{z^+_{ij}}{z^+_j} - \beta\frac{z^-_{ij}}{z^-_j}\right) R^j_{l+1}. \tag{2}$$

Here, $z^{+/-}_{ij}$ refers to the amount of positive/negative input that node i contributed to node j. The individual contributions are divided by the sum over all positive/negative contributions of the nodes in layer l, $z^{+/-}_j = \sum_i z^{+/-}_{ij}$, such that the relevance is conserved from layer $l+1$ to layer l. We have chosen this rule, as it allows for adjusting how much weight is put on positive contributions relative to inhibitory contributions that benefit the AD score. LRP with a β value of zero allows only positive contributions to be shown in the heatmap, whereas non-zero β values additionally correct for the inhibitory effects of neuron activations. When diagnosing AD, the network needs to balance structural evidence speaking for and against AD. Any given local area that looks *healthy* to the network, might have inhibitory effects on the AD score, as it correlates more with HC patients.

As the network increases its receptive field size throughout the layers, healthy areas within this receptive field might inhibit the contribution of affected areas to the final class score of AD. By reversing this process with LRP, positive contributions lying closer to healthy areas will thus obtain a lower relevance score, as they overlap with inhibited receptive fields. This leads to sparser heatmaps, see also Binder et al. (2016a), and might disproportionately affect small structures surrounded by "healthy areas." As AD—especially in the early stages of the disease—can affect brain areas in a highly localized manner, heatmaps obtained with lower β values might therefore be more meaningful, as they highlight *all* positive contributions, irrespective of their surroundings. Accordingly, we focus in the present study on $\beta = 0$, but additionally test the robustness for varying values of β ($\beta = 0, 0.25, 0.5, 0.75, 1$).

For a more detailed description of the LRP algorithm, we kindly refer the reader to Bach et al. (2015); Montavon et al. (2018). A PyTorch implementation of the LRP algorithm has been developed for the current work and is available on github[4].

2.3.2. Guided Backpropagation (GB)

In order to emphasize and point out the advantages of LRP as a diagnostic tool, we compared it to a gradient-based method, the guided backpropagation (GB) algorithm (Springenberg et al., 2014). In GB, the absolute values of the gradient of the output with respect to the input nodes is shown as a heatmap, with the additional twist that negative gradients are set to zero at the rectification layers of the network. As was shown by Rieke et al. (2018), "rectifying" the gradients in the backward pass leads to more focused heatmaps.

2.4. Analyzing the Classification Decisions

The CNN model was evaluated on each MR image from the test set and, subsequently, both the LRP as well as the GB algorithm were used to produce a heatmap for each MR image. In the case of LRP, we produced separate heatmaps for each β value. We analyzed the resulting heatmaps (1) group-wise to distill those regions, which are particularly "important" for the AD classification and (2) individually to understand the network decisions per sample and find differences between subjects. For the former, we computed an average AD heatmap (obtained from all AD subjects) and an average HC heatmap (obtained from all HCs), which we then further split into a true positive heatmap (i.e., average heatmap of clinically validated AD patients, who are classified as AD), a false positive heatmap (i.e., average heatmap of HCs classified as AD), a true negative heatmap (i.e., average heatmap of HCs classified as HC) and a false negative heatmap (i.e., average heatmap of clinically validated AD patients classified as HC). For GB, these heatmaps highlight those areas to which the network is on average most susceptible. For LRP, they show the average relevance of each voxel for contributing to the AD score. All LRP heatmaps show the average relevance for the same class (AD), such that they can be compared on the same scale (relevance for AD diagnosis). As the AD scores of HCs typically range between 0 and 0.5, there will be relevance for AD in HCs, too.

[4]https://github.com/moboehle/Pytorch-LRP

2.5. Atlas-Based Importance Metrics

To quantitatively analyze the heatmaps and the underlying CNN model, we assessed the importance of different brain areas—as defined by the Neuromorphometrics brain atlas (Bakker et al., 2015)—by using the following three metrics for both LRP and GB.

2.5.1. Sum of AD Importance per Area

As a first metric of importance, the resulting heatmap values were simply summed per area. While this can already be taken as a measure of importance, the resulting importance scores are highly correlated to the area size, see **Figure 4**. Therefore, two size-independent metrics for importance were additionally analyzed in more detail: the size-normalized sum, and the average gain (ratio) when comparing to the average HC patient.

2.5.2. Size-Normalized AD Importance Metric

For diagnostic purposes, it can be particularly interesting to identify areas that over their entire volume carry a lot of information, i.e., areas with high *relevance density* or, in GB, *susceptibility density*. Therefore, we divided here the sum of AD importance per area by the size of the area (i.e., number of voxels), which corresponds to the regional mean relevance/susceptibility. While low values over large areas might be due to statistical fluctuations in the data, clusters of relevance (LRP) or susceptibility (GB) in a very confined area could be indicative of the presence of certain biomarkers for AD.

2.5.3. Gain—Ratio of Values With Respect to the Average HC

Lastly, it is important to note that HCs are not "relevance-free" under the LRP algorithm: HCs might exhibit certain structural elements in their brains that are correlated with the AD diagnosis. While the network might still classify them as HC, these structures lead to a class score greater than zero for virtually every subject. Thus, as an additional metric, we will look at the "gain" in relevance (LRP) and susceptibility (GB) per area, i.e., the ratio to the average HC in that area. By doing this, those areas that differ most between the two cases will be attributed the highest importance.

3. RESULTS

In section 3.1, we compare the heatmaps generated by GB and LRP qualitatively with respect to different β values and different sets of data (AD, HC, true positives, false positives etc., see **Figures 2**, **3**). In section 3.2, we quantitatively compare the heatmaps with respect to the different atlas-based importance metrics (see **Figures 4**–**7**). In section 3.3, we present and discuss the LRP heatmaps of two individual patients (see **Figure 8**) and investigate the association between LRP relevance scores and hippocampal volume as one of the neurobiological key markers of AD (see **Figure 9**).

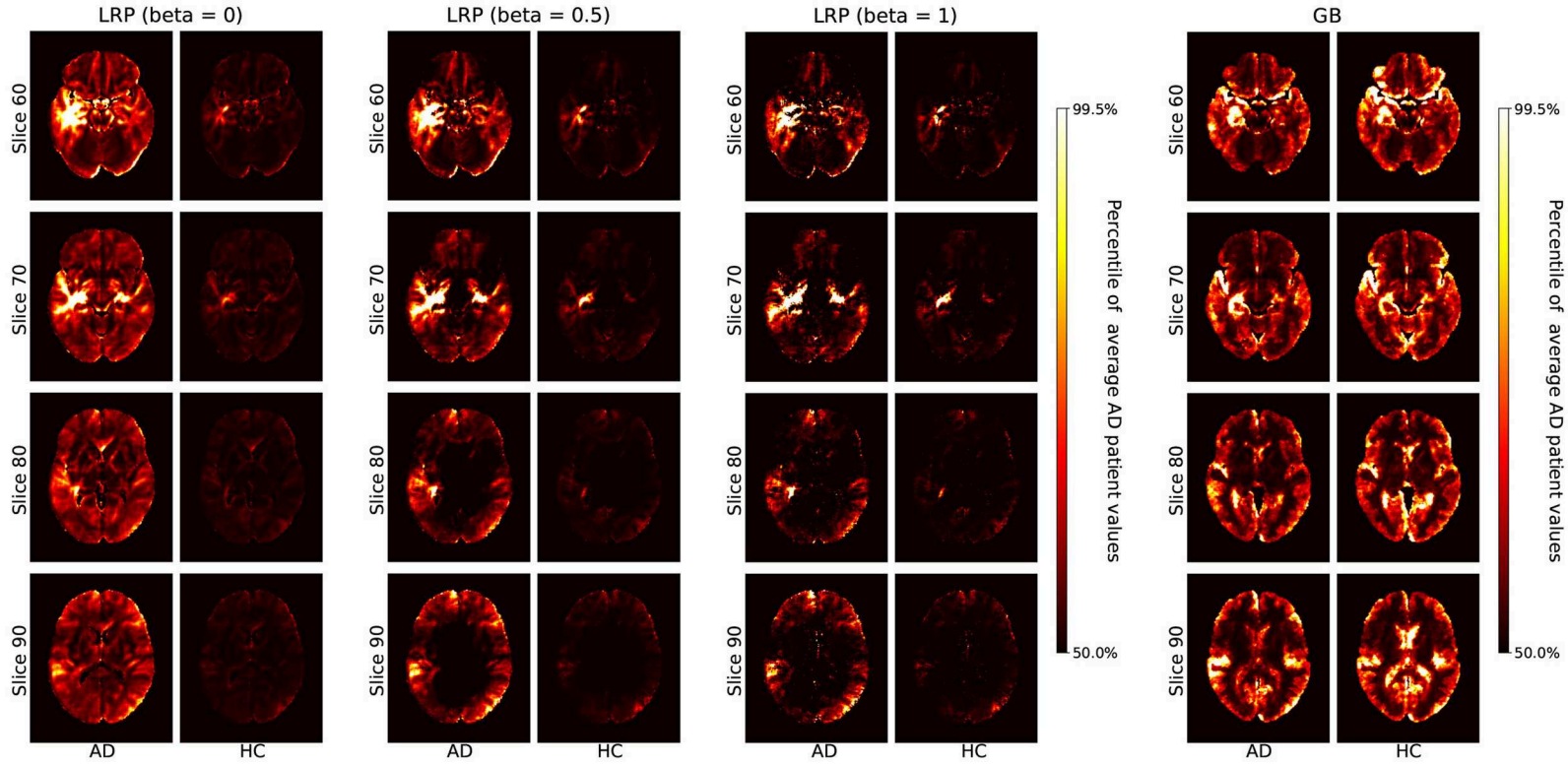

FIGURE 2 | Average heatmaps for AD patients and healthy controls (HCs) in the test set are shown separately for LRP with $\beta = 0, 0.5, 1$ **(Left)** and GB **(Right)**. The scale for the heatmap is chosen relative to the average AD patient heatmap for LRP and GB respectively. Hence, values in the average heatmaps that are higher than the 50th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are black (white).

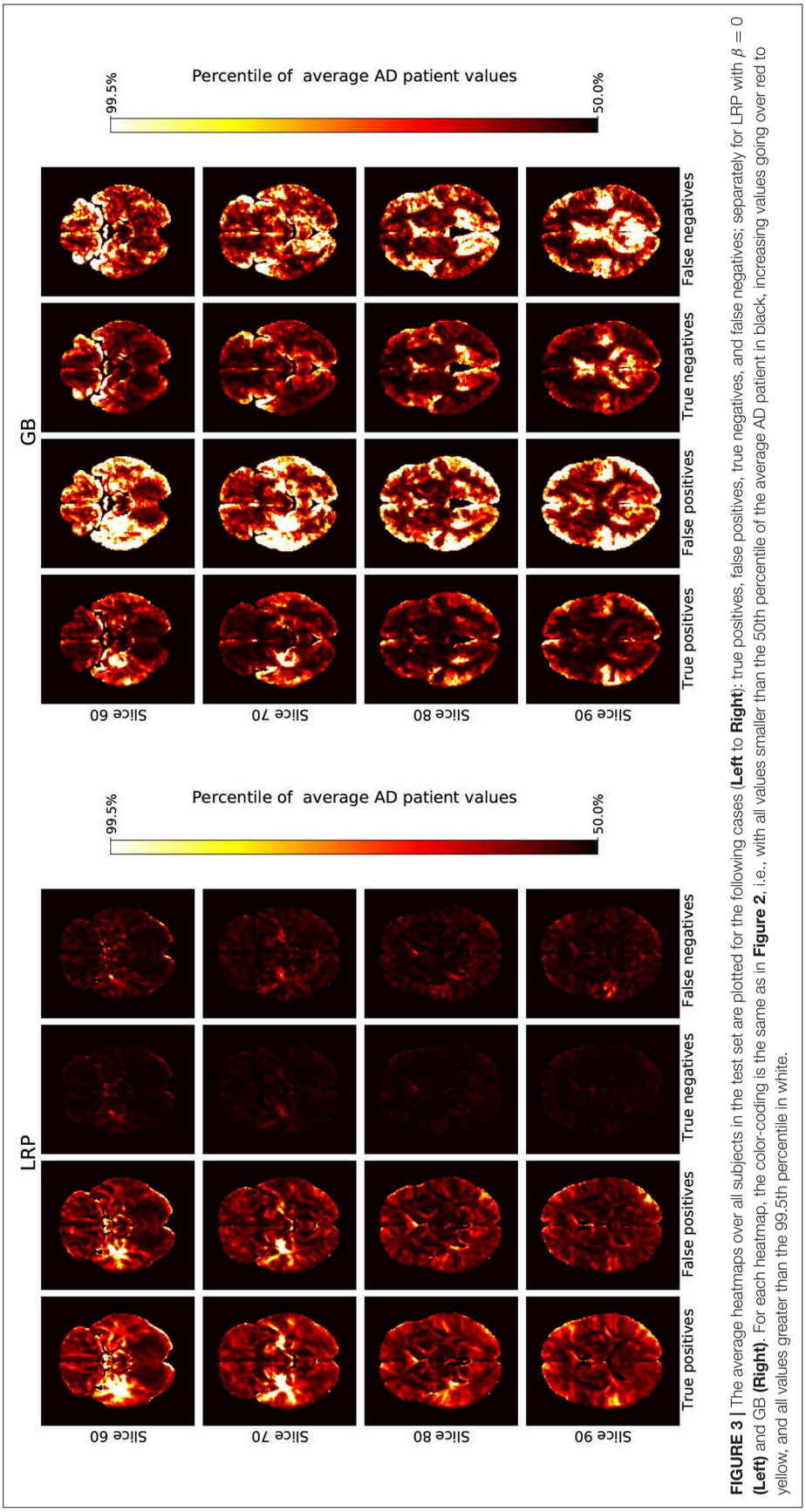

FIGURE 3 | The average heatmaps over all subjects in the test set are plotted for the following cases (**Left** to **Right**): true positives, false positives, true negatives, and false negatives; separately for LRP with $\beta = 0$ **(Left)** and GB **(Right)**. For each heatmap, the color-coding is the same as in **Figure 2**, i.e., with all values smaller than the 50th percentile of the average AD patient in black, increasing values going over red to yellow, and all values greater than the 99.5th percentile in white.

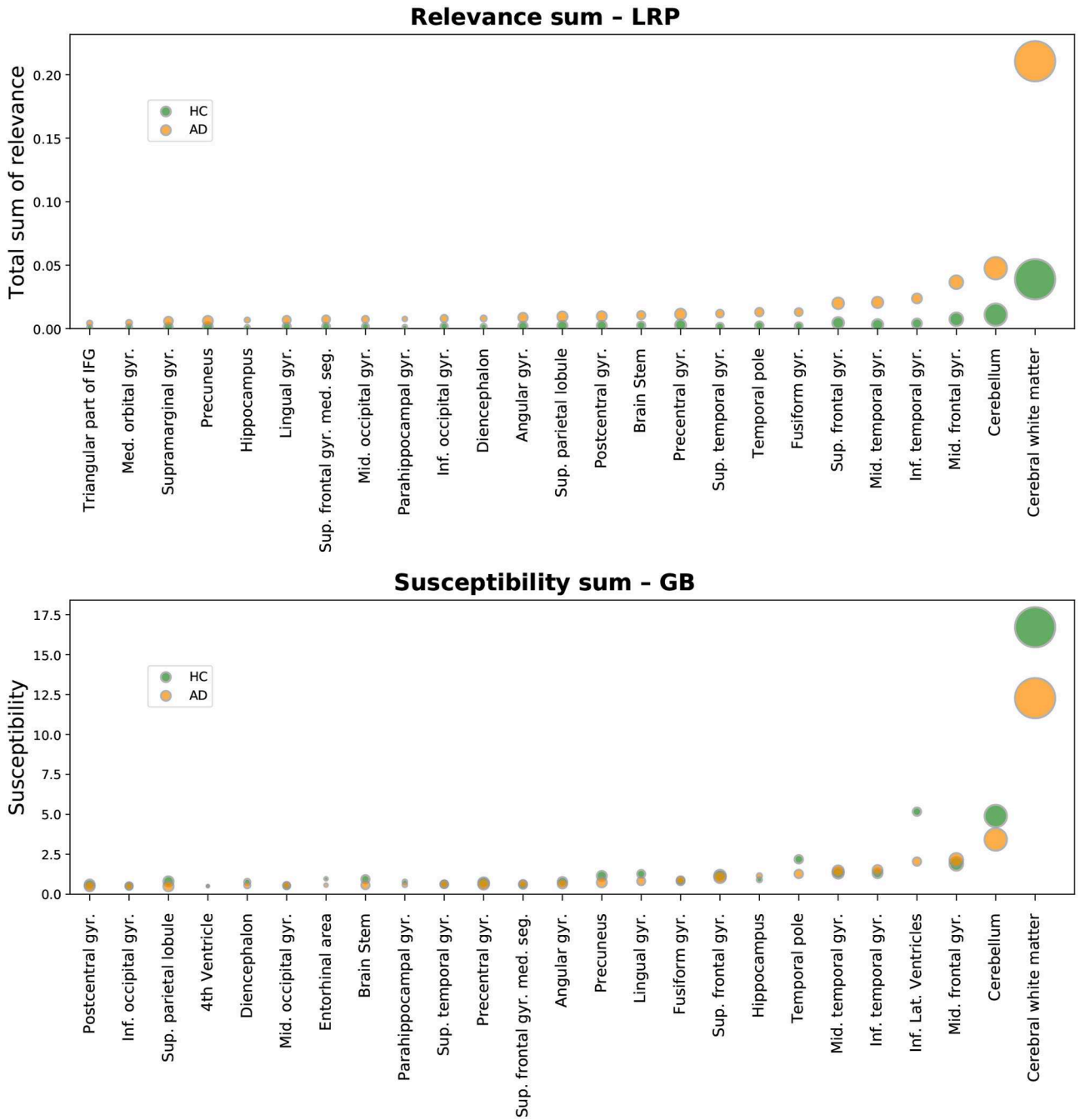

FIGURE 4 | Absolute sum of relevance (LRP, **Top**) and absolute sum of susceptibility (GB, **Bottom**) is shown for different brain areas. Susceptibility refers to the absolute value of the GB gradients. Only the top 25 most important areas under this metric are shown for LRP and GB respectively. The circles show the average sum for each area over all AD patients (orange) and all healthy controls (HCs, green) in the test set. By setting the metric to linearly scale with the corresponding brain area size, it becomes clear that this metric is correlated with the size of the brain areas.

3.1. Average Heatmap Comparison

In **Figure 2**, we show the average heatmaps for AD patients and HCs, separately for LRP with different β values ($\beta = 0, 0.5, 1$) and GB. The AD pattern between LRP and GB is relatively similar, which is reasonable because all heatmaps are extracted from the same CNN model. However, whereas GB heatmaps are very susceptible for both AD and HCs, LRP heatmaps show much more relevance in AD patients than HCs. This indicates that LRP heatmaps might be more valuable in assessing why a certain person has been classified as AD patient as opposed

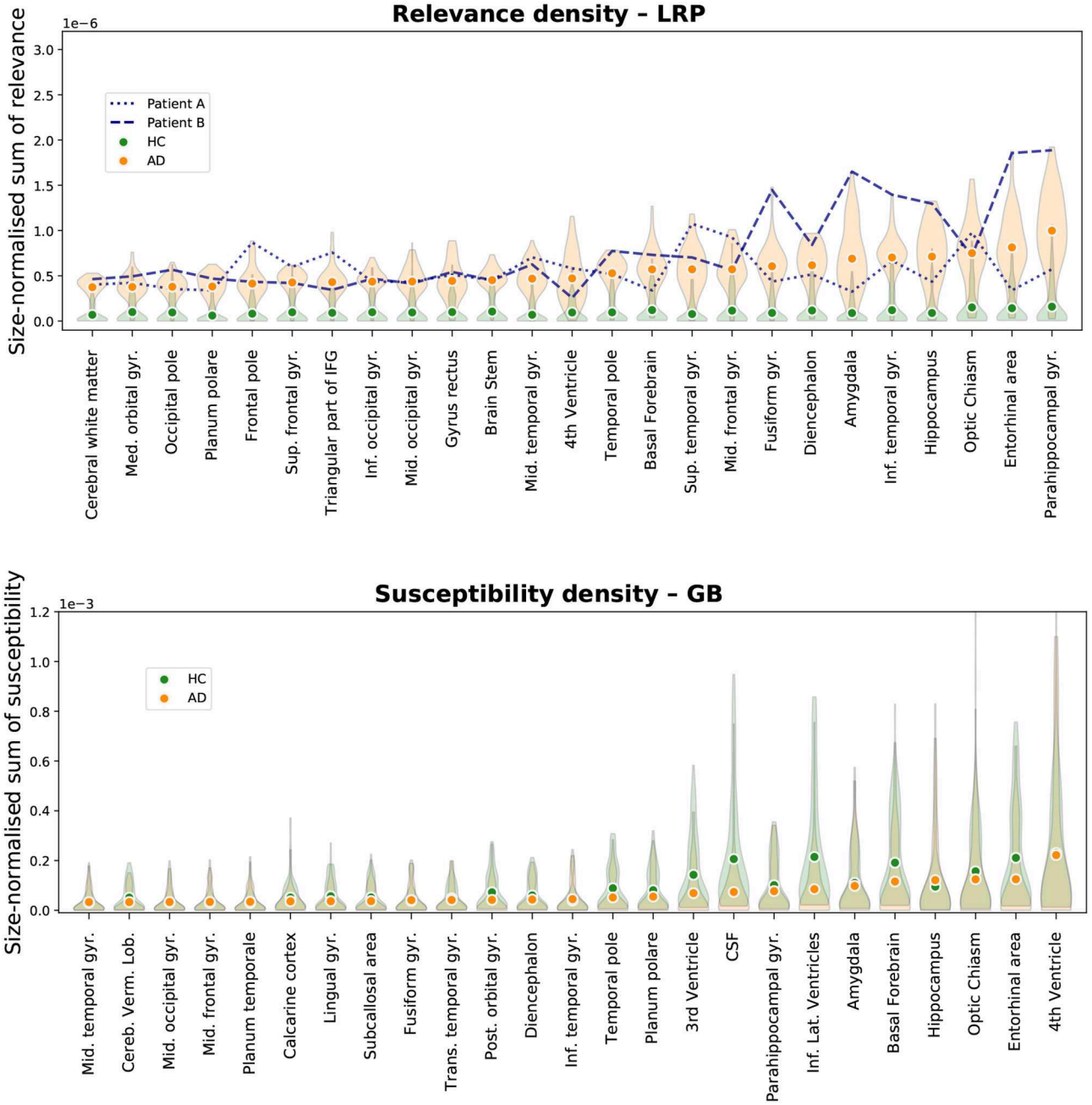

FIGURE 5 | Size-normalized relevance (LRP, **Top**) and size-normalized susceptibility (GB, **Bottom**) is shown for different brain areas. Only the top 25 most important areas under this metric are shown for LRP and GB respectively. We show the average density for all AD patients (orange circles) and all healthy controls (HCs, green circles) in the test set along with a density estimation of the distribution of values per area (orange and green shaded area for AD and HCs respectively). Moreover, two patients were selected to emphasize the diversity in relevance distributions for LRP; the patients were selected as those with the highest cosine distance in the relevance-density space of the 25 areas between each other among those patients that were classified as AD with a class score >90%.

to which voxels should be changed to increase the likelihood for AD diagnosis. Concerning the different β values, it is noted that the heatmaps look qualitatively similar, but that sparseness increases with higher β values (which is due to a larger effect of inhibitory contributions, see also Binder et al., 2016a). Since β values close to 0 focus on positive AD contributions and are thus clinically better interpretable, we focus on $\beta = 0$ in the remaining analyses.

In **Figure 3**, we show the average heatmaps for the distinct classification cases (true positives, false positives etc.), separately for LRP ($\beta = 0$) and GB. In particular, the false positives lead to an interesting insight: For LRP, the false positives

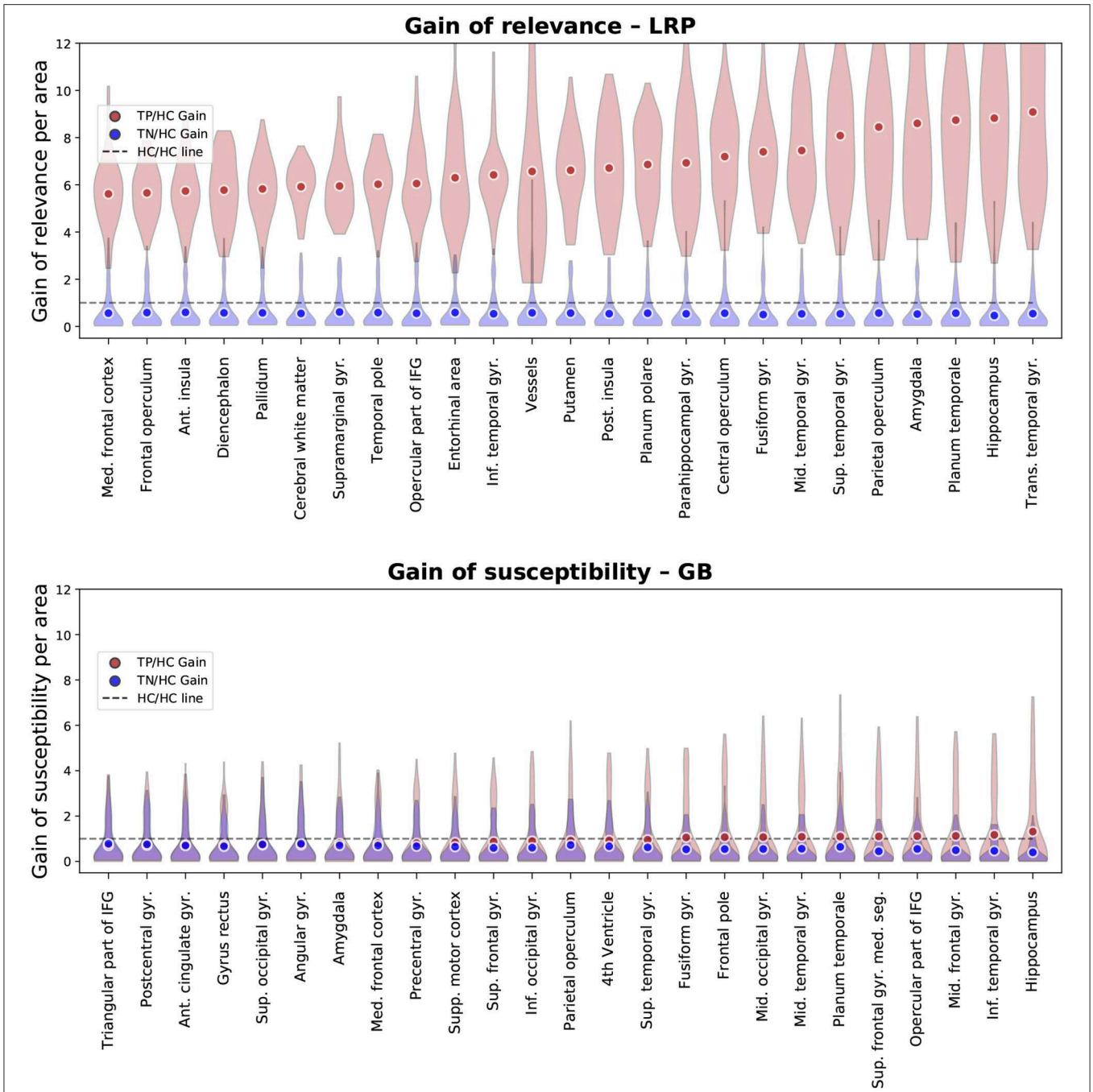

FIGURE 6 | Gain ofrelevance (LRP, **Top**) and gain of susceptibility (GB, **Bottom**) is shown for different brain areas. The gain per area is defined as the average sum of relevance (LRP) or susceptibility (GB) in a given area divided by the average sum in this area over all healthy controls (HCs) in the test set. Again, only the top 25 most important areas under this metric are shown for LRP and GB respectively. To provide an estimate of gain in correctly classified subjects, we show here the mean and density estimations only for true positive (TP) and true negative (TN) classifications. As an additional visual aid, the identity gain (gain of 1) is shown as a dashed line.

exhibit less relevance than the true positives, but generally in similar areas. This could indicate that in these patients structures that are correlated with AD were found, albeit that overall the positive contribution was less compelling than for true AD patients. For GB, on the other hand, the false classifications (mostly false positives, but also false negatives) seem to exhibit the highest gradient values of all cases. This exemplifies well what GB truly measures: in the case of false positives (and negatives), the network might be "unsure" and more easily influenced to change its decision; the outcome is unstable. The highlighted areas that could change the outcome are very broadly distributed and need

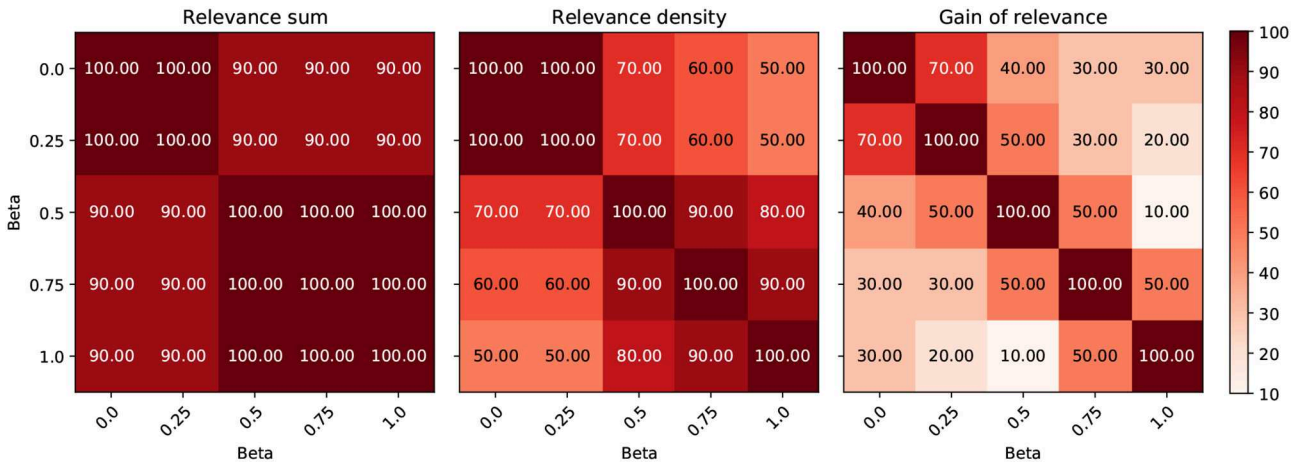

FIGURE 7 | Comparison of the effect of different β values on the regional ordering in **Figures 4–6**. The intersection between the top 10 regions of the three metrics is shown for different LRP β values in %.

not necessarily represent areas with positive contributions for AD.

3.2. Atlas-Based Importance Metrics

In **Figure 4**, we show the sum of AD importance per area, separately for LRP ($\beta = 0$) and GB. Although this metric seems to be dominated by the size of the respective brain area, one important qualitative difference between LRP and GB is visible: in the LRP results, the mean importance values per area are consistently much higher for AD patients than for HCs. For GB, this clear split is not present; moreover, the average sum of gradients in several brain regions, including the cerebral white matter and cerebellum, is even higher for HC than for AD. This exemplifies well that the heatmaps for GB cannot directly be interpreted as showing the relevance for AD classification, but instead show the sensitivity of the outcome to certain areas, which does not have to be AD or HC specific. As the absolute sum of importance correlates with the size of the respective brain area, the following metrics, in which we controlled for the brain area size, might be better interpretable.

In **Figure 5**, the total sum of importance is normalized by the size of the respective brain area. Here, the aforementioned difference in the distributions between HCs and AD patients becomes even more apparent: while the distributions are very heavily overlapping for GB, this is not the case for LRP. Notably, the variance in the AD distributions is much higher in the AD case than in the HC case. This could indicate that the network has learned to differentiate between subtypes of AD and bases its decision on different structural elements for different patients; the existence of different subtypes of AD has been investigated in recent work, see for example (Ferreira et al., 2017; Park et al., 2017). In contrast, for HCs the relevance density is consistently very low. As an example of the diversity in importance assessments according to this metric, we added the "individual fingerprints" of two AD patients to **Figure 5**; for these patients the individual heatmaps will be compared in section 3.3 and **Figure 8**.

In **Figure 6**, the results for the gain metric for different cases—true positives and true negatives—are visualized. This metric allows for plotting the LRP and the GB results on the same scale and emphasizes once again the stronger distinction between AD patients and HCs under the LRP algorithm. Most gain for LRP has been found in areas of temporal lobe including transversal temporal gyrus, hippocampus, planum temporale and amygdala.

In **Figure 7**, we compare the regional overlap of the top 10 regions between the β values $0, 0.25, 0.5, 0.75, 1$, separately for the three importance metrics. It is shown that (1) the regional overlap is strongest for relevance sum followed by relevance density and relatively unstable for gain of relevance especially for large and more distant β values and (2) the regional overlap is–as expected–stronger for neighboring β values. The instability of the gain metric for higher β values is probably due to the associated sparsity leading to very low relevance scores for HCs (which might—in some cases—inflate the gain metric).

3.3. Individual Heatmaps—Fingerprinting and Neurobiological Relevance

Since the LRP heatmaps take into account the individual filter activations and therefore highlight positive contributions to the class score of AD, they might serve as "individual fingerprints" in a diagnostic tool. In **Figure 8**, we show several slices of the relevance heatmaps for two patients in order to highlight the diversity in those heatmaps. The two patients were selected from the test set as those with the highest cosine distance in the relevance-density space between each other among those patients that were classified as AD with a class score $> 90\%$ (their individual trajectories of region-wise relevance are shown in **Figure 5**). It can be seen that the areas, which mainly contributed to the network decision, are rather different for the two patients. For one patient (patient B), the class score of the network is heavily influenced by areas of the temporal lobe, such as

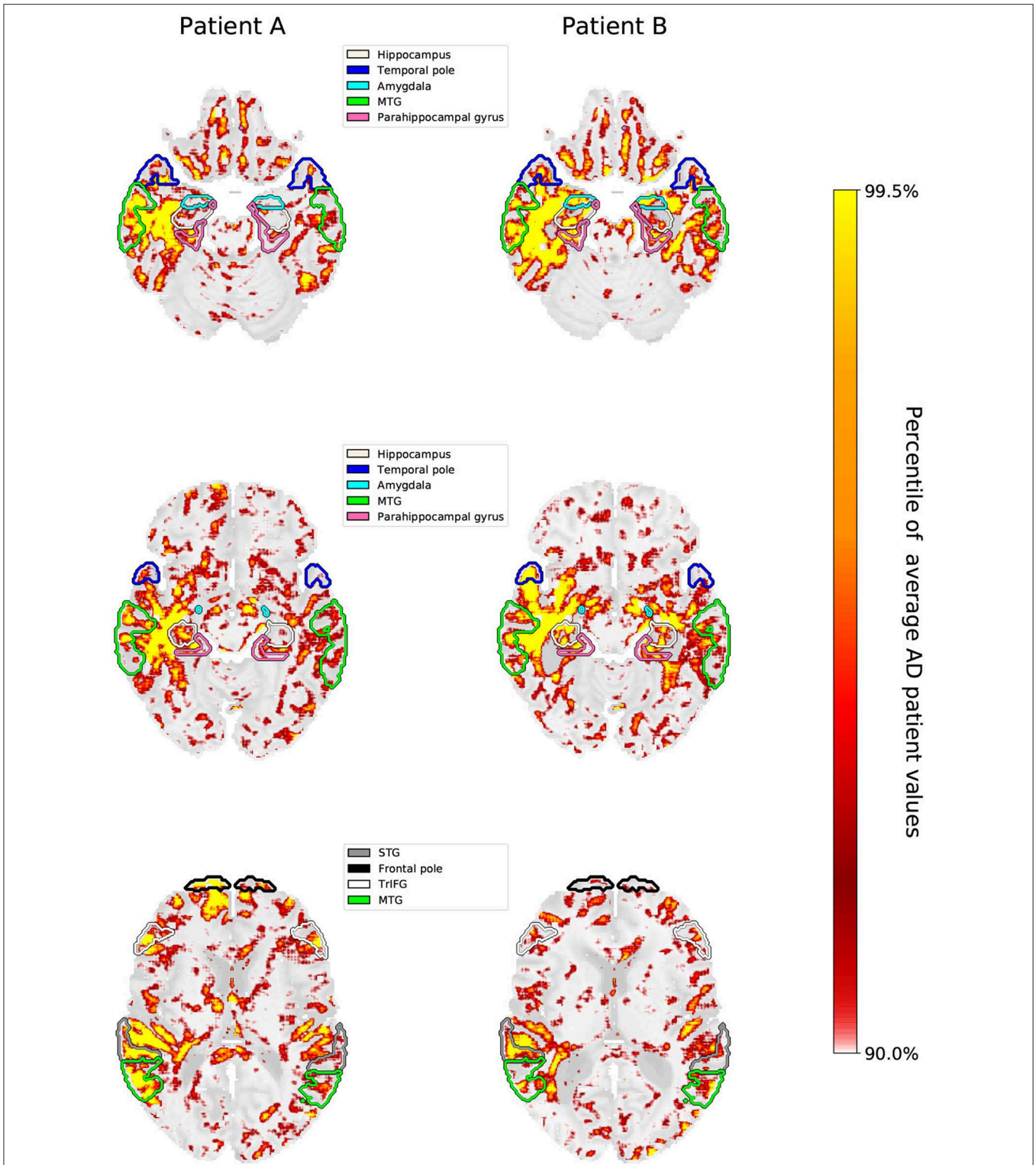

FIGURE 8 | Three brain slices are shown for patient A and patient B, whose individual slopes in relevance density have been shown in **Figure 5**. The highlighted areas are the hippocampus, temporal pole, amygdala, parahippocampal gyrus, medial temporal gyrus (MTG), superior temporal gyrus (STG), triangular part of the inferior frontal gyrus (TrIFG) and frontal pole. The scale for the heatmap is chosen relative to the average AD patient heatmap. Hence, values in the individual patients that are higher than the 90th percentile and lower than the 99.5th percentile are linearly color-coded as shown on the scale. Values below (above) these numbers are transparent (yellow).

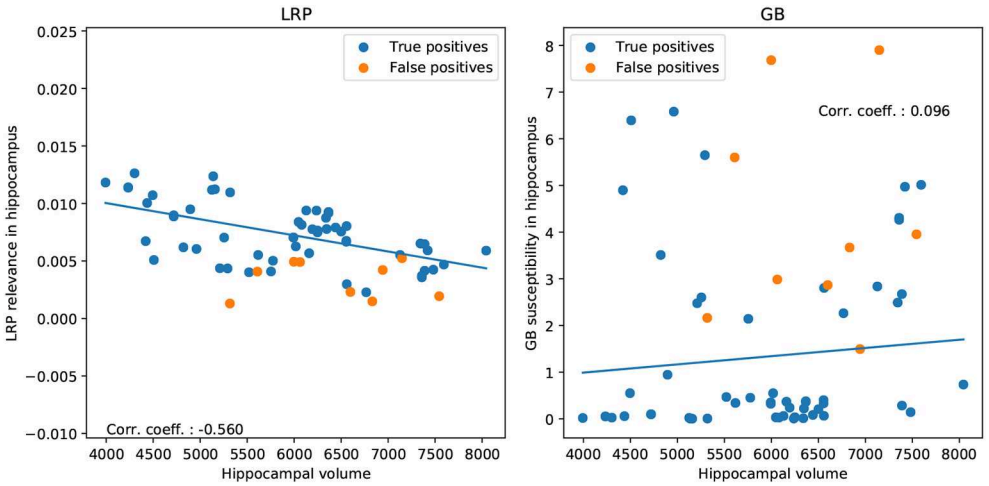

FIGURE 9 | Correlation between hippocampal volume and LRP relevance/GB susceptibility in hippocampus for correctly classified AD patients (true positives; **Left**: LRP, **Right**: GB). For illustration, we show additionally the false positive classifications.

parahippocampal gyrus, entorhinal area, hippocampus, inferior temporal gyrus and amygdala, while for the second patient (patient A), frontal areas, including triangular part of inferior frontal gyrus, superior frontal gyrus and frontal pole, in addition to superior temporal gyrus seem to be most informative.

To investigate whether higher importance scores correspond to stronger anatomical deviations (e.g., atrophy) in correctly classified AD patients (true positives), we performed a correlation analysis between hippocampal volume and LRP relevance/GB susceptibility scores (see **Figure 9**). We show that the LRP relevance score ($\beta = 0$) in the hippocampus is significantly (negatively) correlated with hippocampal volume ($-0.560, p < 10^{-3}$, permutation test), whereas the GB score is not ($0.096, p = 0.77$). To rule out that false positives are outliers in terms of association between hippocampal volume and LRP relevance, we included them in **Figure 9**. Interestingly, for larger β values the correlation tends to decrease (-0.560, -0.562, -0.525, -0.457, -0.361 for $\beta = 0, 0.25, 0.5, 0.75, 1$ respectively) supporting our notion of a higher neurobiological relevance in case of β values close to 0.

4. DISCUSSION

In this study, we introduced LRP as a powerful method for explaining individual CNN decisions in AD classification. After training a CNN to separate AD patients and HCs based on structural MRI data, individual heatmaps—indicating the importance for each voxel for the respective classification decision—were produced for the test subjects. We analyzed the heatmaps with respect to different classification subgroups (AD patients, HCs, true positives, false positives etc.) and different β values. The relevance of brain regions contained in the Neuromorphometrics atlas was evaluated using three different importance metrics, namely the sum of importance per area, the size-normalized AD importance, and the gain as ratio between AD and HC importance. We demonstrated that LRP-derived heatmaps—in contrast to GB—provide (1) high specificity for individuals and (2) little relevance for AD in HCs. Additionally, areas that exhibit a lot of relevance correlate well with what is known from literature. Importantly, these LRP heatmaps were produced without the need for expert annotations on the presence or absence of biomarkers throughout the learning process. This combination of a simple classification task (AD vs. HC) and in-depth network analysis by LRP might be a promising tool for diagnostics. Additionally, it could allow for discovering new and unknown biomarkers for a variety of diseases and might help distinguishing subtypes of AD by analyzing the diversity in "relevance hot-spots" across all AD patients. Furthermore, the size-corrected metrics ("relevance density" and "relevance gain") seem to correlate well with what is known from AD research, indicating that the most discriminating features for classifying an input image as AD can be found in the temporal lobe. We therefore think that a well-trained neural network, analyzed by means of the LRP algorithm, can become a useful tool for practitioners and increase the trust in computer-aided diagnoses, as an interpretable explanation of the decision can be produced.

4.1. Regional Specificity of LRP

We quantitatively evaluated the heatmaps, obtained by either GB or LRP, toward different brain areas according to the Neuromorphometrics atlas (Bakker et al., 2015) by summarizing the importance (AD relevance in case of LRP, susceptibility in case of GB) for each brain area separately. Both types of heatmaps mostly identified regions known to be important in disease progression of AD, such as structures in the medial temporal lobe including hippocampus, amygdala, parahippocampal gyrus, and entorhinal cortex (Du et al., 2001; Desikan et al., 2009; Frisoni et al., 2010; Velayudhan et al., 2013; Weiner et al., 2013; Klein-Koerkamp et al., 2014; Long et al., 2017) as well as frontal and parietal areas (Casanova et al., 2011; Quiroz et al., 2013; Kilimann et al., 2017; Park et al., 2017; Liu et al., 2018). For

all these regions morphometric changes including global and local atrophy (e.g., smaller volumes of hippocampus or amygdala, reduced cortical thickness or gray matter density) or deviations in shape have been shown and related to disease progression and cognitive decline (Desikan et al., 2009; Frisoni et al., 2010; Weiner et al., 2013; Hidalgo-Muñoz et al., 2014; Long et al., 2017; Ledig et al., 2018). These changes seem to be utilized by our CNN framework for making individual predictions and are highlighted in the heatmaps of both LRP and GB. However, the contrast in importance scores between AD patients and HCs is much higher for LRP than GB (in GB, the average heatmaps for AD patients and HCs are quite similar). This supports the notion that LRP heatmaps reflect AD-specific relevance, whereas GB emphasizes areas which the network more generally is sensitive to. Regarding other structures found to be important in our network, it might be interesting to see if also other neural networks find relevance in these areas and if predictions about finding significant structural changes in these areas might be possible at some point. In this respect, the decisions of such networks can be treated as a "second opinion" and a reciprocal learning process with medical experts might be initiated.

4.2. Fingerprinting and Neurobiological Relevance

In addition to heatmap differences between AD patients and HCs, we noticed a high variability between the heatmaps of individual AD patients for the LRP method. This variability was not only reflected in a high variance of important scores within regions, but also in individual trajectories ("fingerprints"), which we exemplary depicted for two AD patients, see **Figure 8**. For future work, it might be very interesting to see if these fingerprints reflect different disease stages of AD (Braak and Braak, 1991; Casanova et al., 2011) or allow for identifying subtypes of AD, in which brain areas are affected differently (Murray et al., 2011; Noh et al., 2014; Scheltens et al., 2016; Zhang et al., 2016; Ferreira et al., 2017; Park et al., 2017). Zhang et al. (2016), for example, identified a temporal, a subcortical and a cortical atrophy factor associated with impairment in different cognitive domains. Another important question is whether the relevance found by the LRP method reflect some true evidence in the sense of biomarkers. By showing that the hippocampal volume is significantly correlated to the LRP relevance scores (but not to the GB susceptibility scores), we argue that LRP—at least partially—succeeded here in breaking down the relevance to the level of voxels in a meaningful way. Interestingly, we found higher correlations for lower β values speaking for a higher neurobiological relevance of β values close to 0. Further studies are needed to more carefully relate LRP relevance measures to other clinical markers of AD including biomarkers and neuropsychological test scores, also in dependency of different CNN models and parameter settings. Moreover, our metrics should be evaluated in patients with mild cognitive impairment (MCI).

4.3. Related Work

Visualization of deep neural networks is a fairly new research area and different attempts have been made to provide intuitive explanations for neural network decisions. However, there is not yet a state-of-the art visualization method as saliency maps for example have been shown to be misleading (Adebayo et al., 2018). In Alzheimer's research, there are only a couple of studies that looked into different visualization methods based on MRI and/or PET data. Most of these studies either visualized filters and activations of the first or last layer (Sarraf and Tofighi, 2016; Lu et al., 2018; Ding et al., 2019) or used the occlusion method to exclude some parts (e.g., with a black patch) of the input image and recalculate the classifier output (Korolev et al., 2017; Esmaeilzadeh et al., 2018; Liu et al., 2018). Based on visual impression, they found that the networks focus primarily on areas known to be involved in AD, such as hippocampus, amygdala or ventricles, but occasionally also other areas such as thalamus or parietal lobe appear. Importantly, in contrast to our study, they did not quantitatively analyze the data, e.g., with respect to brain areas contained in an atlas or underlying neurobiological markers. Additionally, they did not compare different visualization methods or looked for inter-individual differences. One study, however, used gradient-weighted classification activation mapping (grad-CAM) and compared it to sensitivity analysis for AD classification (Yang et al., 2018). They demonstrate that these different visualization methods capture different aspects of the data and show high variability depending e.g., on the resolution of the convolutional layers. In Rieke et al. (2018), gradient-based and occlusion methods (standard patch occlusion and brain area occlusion) were qualitatively and quantitatively compared for AD classification. High regional overlaps between the methods, mostly inferior and middle temporal gyrus, were found but for gradient-based methods the importance was more widely distributed. Regarding the LRP method, we are only aware of one application in the neuroimaging field: Thomas et al. (2018) introduce interpretable recurrent networks for decoding cognitive states based on functional MRI data and demonstrate that the LRP method is capable of identifying relevant brain areas for the different tasks and different levels of data granularity.

4.4. Limitations

Although LRP heatmaps seem to be a promising tool for visualizing neural network decisions, we would like to point out several limitations of LRP and other heatmap methods in the context of this study.

First, heatmap methods are limited by the lack of a ground truth. Most commonly, heatmaps are qualitatively evaluated based on visual assessment, but there are also studies proposing sanity checks (Adebayo et al., 2018) or more objective quality measures such as region perturbation (Samek et al., 2015). In Lipton (2018), the interpretability of models has been generally investigated and questioned. In medical research, heatmaps can be qualitatively evaluated based on prior knowledge (e.g., hippocampus is known to be strongly affected in AD, therefore it seems reasonable to find relevance there). Given that in the specific case of heatmaps for MR images the input space is highly structured, we proposed here additional ways for assessing the quality of explanations by using a brain atlas. Future

studies might assess the neurobiological validity by removing presumably important brain areas and re-training the classifier.

Second, it is largely acknowledged that heatmaps are quite sensitive to the specific algorithms (and its parameters, e.g., the β value in case of LRP) used to produce them. However, regarding the β values in LRP, we have shown that the heatmaps are relatively robust toward this parameter, only sparsity increases as a function of β. Additionally, we demonstrated that the regional ordering is relatively stable for relevance sum and density, but unstable for the gain metric—especially in the case of large and more distant β values.

Third, heatmaps just highlight voxels that contributed to a certain classifier decision, but do not allow making a statement about the underlying reasons (e.g., atrophy or shape differences) or potential interactions between voxels or brain areas. For example, it is difficult to disentangle interactions between different regions (certain patterns in the hippocampus might only be considered as positive evidence if structure Y is found in area Z) nor do we know whether the network developed specific filters for atrophy or the shapes of different structures. Although we found in this study a significant correlation between hippocampal volume and LRP relevance measures, we can not make any claim about causal relationships here. Future studies are necessary to more systematically investigate the relationship between manifested neurobiological markers and LRP explanations.

Fourth, heatmaps strongly depend on the type and quality of the classifier, whose decisions are sought to be explained. Therefore, each heatmap should be read as an indication of where the specific network model sees evidence. For badly trained networks, this does not have to correlate at all with the presence of actual biomarkers. Nevertheless, the better the classifier, the more likely it becomes that the classifier uses meaningful patterns as a basis for its decision and that the heatmaps correlate with "true" evidence for AD. However, heatmaps are also useful in cases, where classification performance is low or sample size is rather small, e.g., for better understanding if the classifier picks up relevant or irrelevant features (e.g., noise or imaging artifacts) and if there are any biases present in the data set (Lapuschkin et al., 2016; Montavon et al., 2018). It would be very interesting to investigate how the heatmaps change for different networks, as those which yield stronger classification results should also base their decisions on *better* "evidence".

And finally, it should be stressed that when we refer to brain areas throughout this work, we refer to the location that the areas are assigned in the brain atlas and not to the individual anatomical structures of any patient. Due to inter-individual differences, the match between the atlas and the individual patient's anatomical realities will not be perfect; this is most likely further aggravated by the presence of atrophy in AD patients.

5. CONCLUSION

In conclusion, we introduced the LRP method for explaining individual CNN decisions in MRI-based AD diagnosis. In contrast to GB, LRP heatmaps can be interpreted as providing individual AD relevance ("What speaks for AD in this particular subject?") as opposed to a general susceptibility for small variations in the input data. Additionally, we provided a framework and specific metrics (i.e., "relevance density" and "relevance gain") to quantitatively compare heatmaps between different groups, brain areas or methods. We demonstrated that these metrics correlate well with clinical findings in AD, but also vary strongly between AD patients. By this, the LRP method might be very useful in a clinical setting for a case-by-case evaluation. However, we would like to point out that (1) our metrics should be evaluated in different network architectures and (2) other (individual) brain atlases might be used for the evaluation of regions. Future studies should evaluate the LRP method on patients with mild-cognitive impairment (MCI) and relate findings to known biomarkers in AD. We are convinced that our framework might also be very useful for other disease classification studies in helping to understand individual network decisions.

DATA AVAILABILITY

The ADNI data set is for researchers publicly available at http://adni.loni.usc.edu/. The code is available at https://github.com/moboehle/Pytorch-LRP.

AUTHOR CONTRIBUTIONS

MB, FE, MW, and KR designed the study. MB and FE engineered the software and analyzed the data. MB, FE, and KR wrote the paper.

FUNDING

We acknowledge support from the German Research Foundation (DFG, 389563835), the Manfred and Ursula-Müller Stiftung and Charité—Universitätsmedizin Berlin (Rahel-Hirsch scholarship and Open Access Publication Fund).

ACKNOWLEDGMENTS

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). Alzheimer's Disease Neuroimaging Initiative was 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.

REFERENCES

Abdel-Hamid, O., Mohamed, A.-R., Jiang, H., Deng, L., Penn, G., and Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Trans. Audio Speech. Language Proc.* 22, 1533–1545. doi: 10.1109/TASLP.2014.2339736

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montréal, QC: Curran Associates, Inc.), 9505–9515.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10:e0130140. doi: 10.1371/journal.pone.0130140

Bakker, R., Tiesinga, P., and Kötter, R. (2015). The scalable brain atlas: instant web-based access to public brain atlases and related content. *Neuroinformatics* 13, 353–366. doi: 10.1007/s12021-014-9258-x

Binder, A., Bach, S., Montavon, G., Müller, K.-R., and Samek, W. (2016a). "Layer-wise relevance propagation for deep neural network architectures," in *Information Science and Applications (ICISA) 2016*, eds K. J. Kim and N. Joukov (Singapore: Springer Singapore), 913–922.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., and Samek, W. (2016b). "Layer-wise relevance propagation for neural networks with local renormalization layers," in *International Conference on Artificial Neural Networks*, eds A. E. P. Villa, P. Masulli, and A. J. P. Rivero (Barcelona: Springer), 63–71.

Bondi, M. W., Edmonds, E. C., and Salmon, D. P. (2017). Alzheimer's disease: past, present, and future. *J. Int. Neuropsychol. Soc.* 23, 818–831. doi: 10.1017/S135561771700100X

Braak, H., and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* 82, 239–259.

Casanova, R., Whitlow, C. T., Wagner, B., Williamson, J., Shumaker, S. A., Maldjian, J. A., et al. (2011). High dimensional classification of structural MRI Alzheimer's Disease data based on large scale regularization. *Front. Neuroinformat.* 5:22. doi: 10.3389/fninf.2011.00022

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature* 538, 20–23. doi: 10.1038/538020a

Desikan, R. S., Cabral, H. J., Hess, C. P., Dillon, W. P., Glastonbury, C. M., Weiner, M. W., et al. (2009). Automated MRI measures identify individuals with mild cognitive impairment and Alzheimer's disease. *Brain* 132, 2048–2057. doi: 10.1093/brain/awp123

Ding, Y., Sohn, J. H., Kawczynski, M. G., Trivedi, H., Harnish, R., Jenkins, N. W., et al. (2019). A deep learning model to predict a diagnosis of Alzheimer disease by using 18 F-FDG PET of the brain. *Radiology* 290, 456–464. doi: 10.1148/radiol.2018180958

Du, A. T., Schuff, N., Amend, D., Laakso, M. P., Hsu, Y. Y., Jagust, W. J., et al. (2001). Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and Alzheimer's disease. *J. Neurol. Neurosurg. Psychiat.* 71, 441–447. doi: 10.1136/jnnp.71.4.441

Esmaeilzadeh, S., Belivanis, D. I., Pohl, K. M., and Adeli, E. (2018). "End to-end Alzheimer's disease diagnosis and biomarker identification," in *International Workshop on Machine Learning in Medical Imaging*, eds Y. ShiHeung-Il and S. Liu (Granada: Springer), 337–345.

Ferreira, D., Verhagen, C., Hernández-Cabrera, J. A., Cavallin, L., Guo, C.-J., Ekman, U., et al. (2017). Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Sci. Rep.* 7:46263. doi: 10.1038/srep46263

Frisoni, G. B., Fox, N. C., Jack C. R. Jr., Scheltens, P., and Thompson, P. M. (2010). The clinical use of structural MRI in Alzheimer disease. *Nat. Rev. Neurol.* 6, 67–77. doi: 10.1038/nrneurol.2009.215

Gupta, A., Ayhan, M., and Maida, A. (2013). "Natural image bases to represent neuroimaging data," in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, Vol 28, eds S. Dasgupta and D. Mcallester (Atlanta, GA: JMLR Workshop and Conference Proceedings), 987–994.

Hidalgo-Muñoz, A. R., Ramírez, J., Górriz, J. M., and Padilla, P. (2014). Regions of interest computed by SVM wrapped method for Alzheimer's disease examination from segmented MRI. *Front. Aging Neurosci.* 6:20. doi: 10.3389/fnagi.2014.00020

Kilimann, I., Hausner, L., Fellgiebel, A., Filippi, M., Würdemann, T. J., Heinsen, H., et al. (2017). Parallel atrophy of cortex and basal forebrain cholinergic system in mild cognitive impairment. *Cereb. Cortex* 27, 1841–1848. doi: 10.1093/cercor/bhw019

Kingma, D. P., and Ba, J. (2015). *Adam: A Method for Stochastic Optimization*. San Diego, CA: ICLR.

Klein-Koerkamp, Y., Heckemann, R., Ramdeen, K., Moreaud, O., Keignart, S., Krainik, A., et al. (2014). Amygdalar atrophy in early Alzheimer's disease. *Curr. Alzheimer Res.* 11, 239–252. doi: 10.2174/1567205011666140131123653

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in Alzheimer's disease. *Brain* 131, 681–689. doi: 10.1093/brain/awm319

Korolev, S., Safiullin, A., Belyaev, M., and Dodonova, Y. (2017). "Residual and plain convolutional neural networks for 3d brain mri classification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (Melbourne, VIC: IEEE), 835–838.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Lake Tahoe, CA: Curran Associates, Inc), 1097–1105.

Lapuschkin, S., Binder, A., Montavon, G., Muller, K.-R., and Samek, W. (2016). "Analyzing classifiers: fisher vectors and deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, CA), 2912–2920.

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed M. A. Arbib (MIT Press), 3361.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

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

Lipton, Z. C. (2018). "The mythos of model interpretability," in *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning*, eds B. Kim, D. M. Malioutov, and K. R. Varshney (New York, NY), 96–100.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Analys.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, M., Cheng, D., Wang, K., and Wang, Y. (2018). Multi-modality cascaded convolutional neural networks for Alzheimer's disease diagnosis. *Neuroinformatics* 16, 295–308. doi: 10.1007/s12021-018-9370-4

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Boston, MA), 3431–3440.

Long, X., Chen, L., Jiang, C., Zhang, L., Initiative, A. D. N., et al. (2017). Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PLoS ONE* 12:e0173372. doi: 10.1371/journal.pone.0173372

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M. F., and Alzheimer's Disease Neuroimaging Initiative. (2018). Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET images. *Sci. Rep.* 8:5697. doi: 10.1038/s41598-018-22871-z

Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Proc.* 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Morris, J. C. (1993). The Clinical Dementia Rating (CDR). *Neurology* 43, 2412–2412–a.

Murray, M. E., Graff-Radford, N. R., Ross, O. A., Petersen, R. C., Duara, R., and Dickson, D. W. (2011). Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol.* 10, 785–796. doi: 10.1016/S1474-4422(11)70156-9

Noh, Y., Jeon, S., Lee, J. M., Seo, S. W., Kim, G. H., Cho, H., Ye, B. S., et al. (2014). Anatomical heterogeneity of Alzheimer disease based on cortical thickness on MRIs. *Neurology* 83, 1936–1944. doi: 10.1212/WNL.0000000000001003

Orrù, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G., and Mechelli, A. (2012). Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci. Biobehav. Rev.* 36, 1140–1152. doi: 10.1016/j.neubiorev.2012.01.004

Park, J.-Y., Na, H. K., Kim, S., Kim, H., Kim, H. J., Seo, S. W., et al. (2017). Robust identification of Alzheimer's disease subtypes based on cortical atrophy patterns. *Sci. Rep.* 7:43270. doi: 10.1038/srep43270

Payan, A., and Montana, G. (2015). Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *CoRR* abs/1502.0.

Quiroz, Y. T., Stern, C. E., Reiman, E. M., Brickhouse, M., Ruiz, A., Sperling, R. A., et al. (2013). Cortical atrophy in presymptomatic Alzheimer's disease presenilin 1 mutation carriers. *J. Neurol. Neurosurg. Psychiat.* 84, 556–561. doi: 10.1136/jnnp-2012-303299

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155, 530–548. doi: 10.1016/j.neuroimage.2017.03.057

Rieke, J., Eitel, F., Weygandt, M., Haynes, J.-D., and Ritter, K. (2018). "Visualizing convolutional networks for MRI-based diagnosis of Alzheimer's disease. in *Understanding and Interpreting Machine Learning in Medical Image Computing Applications* (Granada: Springer), 24–31.

Ritter, K., Lange, C., Weygandt, M., Mäurer, A., Roberts, A., Estrella, M., et al. (2016). Combination of structural MRI and FDG-PET of the brain improves diagnostic accuracy in newly manifested cognitive impairment in geriatric inpatients. *J. Alzheimer's Dis.* 54, 1319–1331. doi: 10.3233/JAD-160380

Ritter, K., Schumacher, J., Weygandt, M., Buchert, R., Allefeld, C., and Haynes, J.-D. (2015). Multimodal prediction of conversion to Alzheimer's disease based on incomplete biomarkers. *Alzheimer's Dement. Diagnos. Assess. Dis. Monitor.* 1, 206–215. doi: 10.1016/j.dadm.2015.01.006

Samek, W., Binder, A., Montavon, G., Bach, S., and Müller, K.-R. (2015). Evaluating the visualization of what a deep neural network has learned. *arXiv* arXiv:1509.06321.

Sarraf, S., and Tofighi, G. (2016). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *bioRxiv*.

Scheltens, N. M., Galindo-Garre, F., Pijnenburg, Y. A., van der Vlies, A. E., Smits, L. L., Koene, T., et al. (2016). The identification of cognitive subtypes in Alzheimer's disease dementia using latent class analysis. *J. Neurol. Neurosurg. Psychiatry* 87, 235–243. doi: 10.1136/jnnp-2014-309582

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv* arXiv:1312.6034.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: the all convolutional net. *arXiv* arXiv:1412.6806.

Sturm, I., Bach, S., Samek, W., and Müller, K.-R. (2016). Interpretable deep neural networks for single-trial EEG classification. *CoRR* abs/1604.0. doi: 10.1016/j.jneumeth.2016.10.008

Suk, H.-I., Lee, S.-W., and Shen, D. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage* 101, 569–582. doi: 10.1016/j.neuroimage.2014.06.077

Thomas, A. W., Heekeren, H. R., Müller, K.-R., and Samek, W. (2018). Interpretable LSTMs for whole-brain neuroimaging analyses. *arXiv* arXiv:1810.09945.

Velayudhan, L., Proitsi, P., Westman, E., Muehlboeck, J.-S., Mecocci, P., Vellas, B., et al. (2013). Entorhinal cortex thickness predicts cognitive decline in Alzheimer's disease. *J. Alzheimer's Dis.* 33, 755–766. doi: 10.3233/JAD-2012-121408

Vieira, S., Pinaya, W. H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: methods and applications. *Neurosci. Biobehav. Rev.* 74, 58–75. doi: 10.1016/j.neubiorev.2017.01.002

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2013). The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's Dement.* 9, e111–e194. doi: 10.1016/j.jalz.2013.05.1769

WHO (2017). *Dementia*. Available online at: https://www.who.int/news-room/fact-sheets/detail/dementia (accessed July 21, 2019).

Yang, C., Rangarajan, A., and Ranka, S. (2018). Visual explanations from deep 3D convolutional neural networks for Alzheimer's disease classification. *arXiv* arXiv:1803.02544.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv* arXiv:1506.06579.

Zeiler, M., and Fergus, R. (2014). "Visualizing and understanding convolutional networks." in *Computer Vision–ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, eds D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars (Zürich: Springer International Publishing), 818–833.

Zhang, X., Mormino, E. C., Sun, N., Sperling, R. A., Sabuncu, M. R., Yeo, B. T. T., et al. (2016). Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc. Natl. Acad. Sci. U. S. A.* 113, E6535–E6544. doi: 10.1073/pnas.1611073113

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Aging Neuroscience

SYSTEMATIC REVIEW
published: 20 August 2019
doi: 10.3389/fnagi.2019.00220

Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data

Taeho Jo [1,2,3]*, *Kwangsik Nho* [1,2,3] and *Andrew J. Saykin* [1,2,3]

[1] Department of Radiology and Imaging Sciences, Center for Neuroimaging, Indiana University School of Medicine, Indianapolis, IN, United States, [2] Indiana Alzheimer Disease Center, Indiana University School of Medicine, Indianapolis, IN, United States, [3] Indiana University Network Science Institute, Bloomington, IN, United States

Deep learning, a state-of-the-art machine learning approach, has shown outstanding performance over traditional machine learning in identifying intricate structures in complex high-dimensional data, especially in the domain of computer vision. The application of deep learning to early detection and automated classification of Alzheimer's disease (AD) has recently gained considerable attention, as rapid progress in neuroimaging techniques has generated large-scale multimodal neuroimaging data. A systematic review of publications using deep learning approaches and neuroimaging data for diagnostic classification of AD was performed. A PubMed and Google Scholar search was used to identify deep learning papers on AD published between January 2013 and July 2018. These papers were reviewed, evaluated, and classified by algorithm and neuroimaging type, and the findings were summarized. Of 16 studies meeting full inclusion criteria, 4 used a combination of deep learning and traditional machine learning approaches, and 12 used only deep learning approaches. The combination of traditional machine learning for classification and stacked auto-encoder (SAE) for feature selection produced accuracies of up to 98.8% for AD classification and 83.7% for prediction of conversion from mild cognitive impairment (MCI), a prodromal stage of AD, to AD. Deep learning approaches, such as convolutional neural network (CNN) or recurrent neural network (RNN), that use neuroimaging data without pre-processing for feature selection have yielded accuracies of up to 96.0% for AD classification and 84.2% for MCI conversion prediction. The best classification performance was obtained when multimodal neuroimaging and fluid biomarkers were combined. Deep learning approaches continue to improve in performance and appear to hold promise for diagnostic classification of AD using multimodal neuroimaging data. AD research that uses deep learning is still evolving, improving performance by incorporating additional hybrid data types, such as—omics data, increasing transparency with explainable approaches that add knowledge of specific disease-related features and mechanisms.

Keywords: artificial intelligence, machine learning, deep learning, classification, Alzheimer's disease, neuroimaging, magnetic resonance imaging, positron emission tomography

OPEN ACCESS

Edited by:
James H. Cole,
King's College London,
United Kingdom

Reviewed by:
Donghuan Lu,
Simon Fraser University, Canada
Zheng Wang,
University of Miami, United States

***Correspondence:**
Taeho Jo
tjo@iu.edu

Received: *01 March 2019*
Accepted: *02 August 2019*
Published: *20 August 2019*

Citation:
Jo T, Nho K and Saykin AJ (2019) Deep Learning in Alzheimer's Disease: Diagnostic Classification and Prognostic Prediction Using Neuroimaging Data. Front. Aging Neurosci. 11:220. doi: 10.3389/fnagi.2019.00220

INTRODUCTION

Alzheimer's disease (AD), the most common form of dementia, is a major challenge for healthcare in the twenty-first century. An estimated 5.5 million people aged 65 and older are living with AD, and AD is the sixth-leading cause of death in the United States. The global cost of managing AD, including medical, social welfare, and salary loss to the patients' families, was $277 billion in 2018 in the United States, heavily impacting the overall economy and stressing the U.S. health care system (Alzheimer's Association, 2018). AD is an irreversible, progressive brain disorder marked by a decline in cognitive functioning with no validated disease modifying treatment (De strooper and Karran, 2016). Thus, a great deal of effort has been made to develop strategies for early detection, especially at pre-symptomatic stages in order to slow or prevent disease progression (Galvin, 2017; Schelke et al., 2018). In particular, advanced neuroimaging techniques, such as magnetic resonance imaging (MRI) and positron emission tomography (PET), have been developed and used to identify AD-related structural and molecular biomarkers (Veitch et al., 2019). Rapid progress in neuroimaging techniques has made it challenging to integrate large-scale, high dimensional multimodal neuroimaging data. Therefore, interest has grown rapidly in computer-aided machine learning approaches for integrative analysis. Well-known pattern analysis methods, such as linear discriminant analysis (LDA), linear program boosting method (LPBM), logistic regression (LR), support vector machine (SVM), and support vector machine-recursive feature elimination (SVM-RFE), have been used and hold promise for early detection of AD and the prediction of AD progression (Rathore et al., 2017).

In order to apply such machine learning algorithms, appropriate architectural design or pre-processing steps must be predefined (Lu and Weng, 2007). Classification studies using machine learning generally require four steps: feature extraction, feature selection, dimensionality reduction, and feature-based classification algorithm selection. These procedures require specialized knowledge and multiple stages of optimization, which may be time-consuming. Reproducibility of these approaches has been an issue (Samper-Gonzalez et al., 2018). For example, in the feature selection process, AD-related features are chosen from various neuroimaging modalities to derive more informative combinatorial measures, which may include mean subcortical volumes, gray matter densities, cortical thickness, brain glucose metabolism, and cerebral amyloid-β accumulation in regions of interest (ROIs), such as the hippocampus (Riedel et al., 2018).

In order to overcome these difficulties, deep learning, an emerging area of machine learning research that uses raw neuroimaging data to generate features through "on-the-fly" learning, is attracting considerable attention in the field of large-scale, high-dimensional medical imaging analysis (Plis et al., 2014). Deep learning methods, such as convolutional neural networks (CNN), have been shown to outperform existing machine learning methods (Lecun et al., 2015).

We systematically reviewed publications where deep learning approaches and neuroimaging data were used for the early detection of AD and the prediction of AD progression. A PubMed and Google Scholar search was used to identify deep learning papers on AD published between January 2013 and July 2018. The papers were reviewed and evaluated, classified by algorithms and neuroimaging types, and the findings were summarized. In addition, we discuss challenges and implications for the application of deep learning to AD research.

DEEP LEARNING METHODS

Deep learning is a subset of machine learning (Lecun et al., 2015), meaning that it learns features through a hierarchical learning process (Bengio, 2009). Deep learning methods for classification or prediction have been applied in various fields, including computer vision (Ciregan et al., 2012; Krizhevsky et al., 2012; Farabet et al., 2013) and natural language processing (Hinton et al., 2012; Mikolov et al., 2013), both of which demonstrate breakthroughs in performance (Boureau et al., 2010; Russakovsky et al., 2015). Because deep learning methods have been reviewed extensively in recent years (Bengio, 2013; Bengio et al., 2013; Schmidhuber, 2015), we focus here on basic concepts of Artificial Neural Networks (ANN) that underlie deep learning (Hinton and Salakhutdinov, 2006). We also discuss architectural layouts of deep learning that have been applied to the task of AD classification and prognostic prediction. ANN is a network of interconnected processing units called artificial neurons that were modeled (Mcculloch and Pitts, 1943) and developed with the concept of Perceptron (Rosenblatt, 1957, 1958), Group Method of Data Handling (GMDH) (Ivakhnenko and Lapa, 1965; Ivakhnenko, 1968, 1971), and the Neocognitron (Fukushima, 1979, 1980). Efficient error functions and gradient computing methods were discussed in these seminal publications, spurred by the demonstrated limitation of the single layer perceptron, which can learn only linearly separable patterns (Minsky and Papert, 1969). Further, the back-propagation procedure, which uses gradient descent, was developed and applied to minimize the error function (Werbos, 1982, 2006; Rumelhart et al., 1986; Lecun et al., 1988).

Gradient Computation

The back-propagation procedure is used to calculate the error between the network output and the expected output. The back propagation calculates the gap repeatedly, changing weights and stopping the calculation when the gap is no longer updated (Rumelhart et al., 1986; Bishop, 1995; Ripley and Hjort, 1996; Schalkoff, 1997). **Figure 1** illustrates the process of the neural network made by multilayer perceptron. After the initial error value is calculated from the given random weight by the least squares method, the weights are updated until the differential value becomes 0. For example, the w_{31} in **Figure 1** is updated by the following formula:

$$w_{31}(t+1) = w_{31}t - \frac{\partial Error Y_{out}}{\partial w_{31}}$$

$$Error Y_{out} = \frac{1}{2}(y_{t1} - y_{o1})^2 + \frac{1}{2}(y_{t2} - y_{o2})^2$$

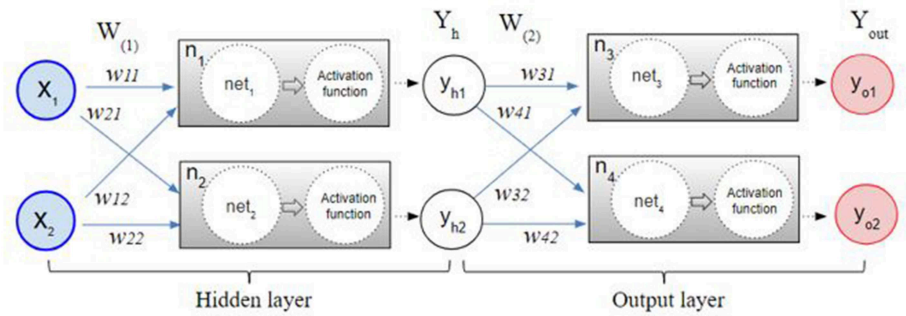

FIGURE 1 | The multilayer perceptron procedure. After the initial error value is calculated from the given random weight by the least squares method, the weights are updated by a back-propagation algorithm until the differential value becomes 0.

The $Error Y_{out}$ is the sum of error y_{o1} and error y_{o2}. y_{t1}, y_{t2} are constants that are known through the given data. The partial derivative of $Error Y_{out}$ with respect to w_{31} can be calculated by the chain rule as follows.

$$\frac{\partial Error Y_{out}}{\partial w_{31}} = \frac{\partial Error Y_{out}}{\partial y_{o1}} \cdot \frac{\partial y_{o1}}{\partial net3} \cdot \frac{\partial net3}{\partial w_{31}}$$

Likewise, w_{11} in the hidden layer is updated by the chain rule as follows.

$$\frac{\partial \text{Error} Y_{out}}{\partial w_{11}} = \frac{\partial \text{Error} Y_{out}}{\partial y_{h1}} \cdot \frac{\partial y_{h1}}{\partial net_1 y} \cdot \frac{\partial net_1}{\partial w_{11}}$$

Detailed calculation of the weights in the backpropagation is described in **Supplement 1**.

Modern Practical Deep Neural Networks

As the back-propagation uses a gradient descent method to calculate the weights of each layer going backwards from the output layer, a vanishing gradient problem occurs as the layer is stacked, where the differential value becomes 0 before finding the optimum value. As shown in **Figure 2A**, when the sigmoid is differentiated, the maximum value is 0.25, which becomes closer to 0 when it continues to multiply. This is called a vanishing gradient issue, a major obstacle of the deep neural network. Considerable research has addressed the challenge of the vanishing gradient (Goodfellow et al., 2016). One of the accomplishments of such an effort is to replace the sigmoid function, an activation function, with several other functions, such as the hyperbolic tangent function, ReLu, and Softplus (Nair and Hinton, 2010; Glorot et al., 2011). The hyperbolic tangent (tanh, **Figure 2B**) function expands the range of derivative values of the sigmoid. The ReLu function (**Figure 2C**), the most used activation function, replaces a value with 0 when the value is <0 and uses the value if the value is >0. As the derivative becomes 1, when the value is larger than 0, it becomes possible to adjust the weights without disappearing up to the first layer through the stacked hidden layers. This simple method allows building multiple layers and accelerates the development of deep learning. The Softplus function (**Figure 2D**) replaces the ReLu function with a gradual descent method when ReLu becomes zero.

While a gradient descent method is used to calculate the weights accurately, it usually requires a large amount of computation time because all of the data needs to be differentiated at each update. Thus, in addition to the activation function, advanced gradient descent methods have been developed to solve speed and accuracy issues. For example, Stochastic Gradient Descent (SGD) uses a subset that is randomly extracted from the entire data for faster and more frequent updates (Bottou, 2010), and it has been extended to Momentum SGD (Sutskever et al., 2013). Currently, one of the most popular gradient descent method is Adaptive Moment Estimation (Adam). Detailed calculation of the optimization methods is described in **Supplement 2**.

Architectures of Deep Learning

Overfitting has also played a major role in the history of deep learning (Schmidhuber, 2015), with efforts being made to solve it at the architectural level. The Restricted Boltzmann Machine (RBM) was one of the first models developed to overcome the overfitting problem (Hinton and Salakhutdinov, 2006). Stacking the RBMs resulted in building deeper structures known as the Deep Boltzmann Machine (DBM) (Salakhutdinov and Larochelle, 2010). The Deep Belief Network (DBN) is a supervised learning method used to connect unsupervised features by extracting data from each stacked layer (Hinton et al., 2006). DBN was found to have a superior performance to other models and is one of the reasons that deep learning has gained popularity (Bengio, 2009). While DBN solves the overfitting problem by reducing the weight initialization using RBM, CNN efficiently reduces the number of model parameters by inserting convolution and pooling layers that lead to a reduction in complexity. Because of its effectiveness, when given enough data, CNN is widely used in the field of visual recognition. **Figure 3** shows the structures of RBM, DBM, DBN, CNN, Auto-Encoders (AE), sparse AE, and stacked AE, respectively. Auto-Encoders (AE) are an unsupervised learning method that make the output value approximate to the input value by using the back-propagation and SGD (Hinton and Zemel, 1994). AE

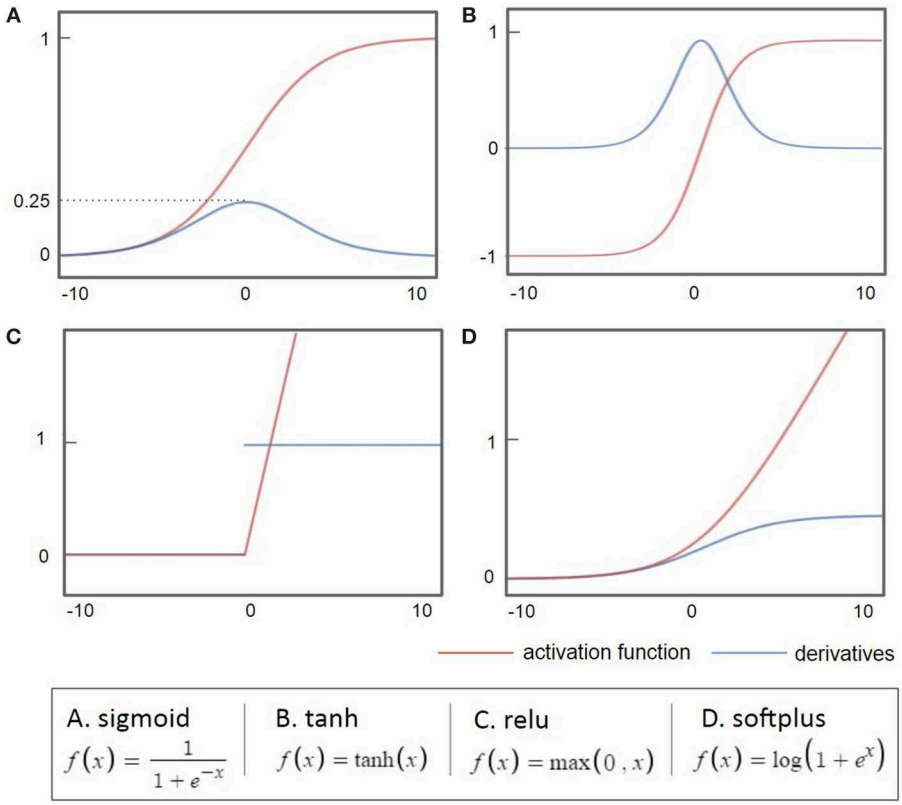

FIGURE 2 | Common activation functions used in deep learning (red) and their derivatives (blue). When the sigmoid is differentiated, the maximum value is 0.25, which becomes closer to 0 when it continues to multiply.

engages the dimensional reduction, but it is difficult to train due to the vanishing gradient issue. Sparse AE has solved this issue by allowing for only a small number of the hidden units (Makhzani and Frey, 2015). Stacked AE stacks sparse AE like DBN.

DNN, RBM, DBM, DBN, AE, Sparse AE, and Stacked AE are deep learning methods that have been used for Alzheimer's disease diagnostic classification to date (see **Table 1** for the definition of acronyms). Each approach has been developed to classify AD patients from cognitively normal controls (CN) or mild cognitive impairment (MCI), which is the prodromal stage of AD. Each approach is used to predict the conversion of MCI to AD using multi-modal neuroimaging data. In this paper, when deep learning is used together with traditional machine learning methods, i.e., SVM as a classifier, it is referred to as a "hybrid method."

MATERIALS AND METHODS

We conducted a systematic review on previous studies that used deep learning approaches for diagnostic classification of AD with multimodal neuroimaging data. The search strategy is outlined in detail using the PRISMA flow diagram (Moher et al., 2009) in **Figure 4**.

Identification

From a total of 389 hits on Google scholar and PubMed search, 16 articles were included in the systematic review.

Google Scholar: We searched using the following key words and yielded 358 results ("Alzheimer disease" OR "Alzheimer's disease"), ("deep learning" OR "deep neural network" OR "CNN" OR "CNN" OR "Autoencoder" OR "DBN" OR "RBM"), ("Neuroimaging" OR "MRI" OR "multimodal").

PubMed: The keywords used in the Google Scholar search were reused for the search in PubMed, and yielded 31 search results ("Alzheimer disease" OR "Alzheimer's disease") AND ("deep learning" OR "deep neural network" OR "CNN" OR "recurrent neural network" OR "Auto-Encoder" OR "Auto Encoder" OR "RBM" OR "DBN" OR "Generative Adversarial Network" OR "Reinforcement Learning" OR "Long Short Term Memory" OR "Gated Recurrent Units") AND ("Neuroimaging" OR "MRI" OR "multimodal").

Among the 389 relevant records, 25 overlapping records were removed.

Screening Based on Article Type

We first excluded 38 survey papers, 22 theses, 19 Preprint, 34 book chapters, 20 conference abstract, 13 none English papers, 5 citations, and 10 patents. We also excluded 11 papers of which

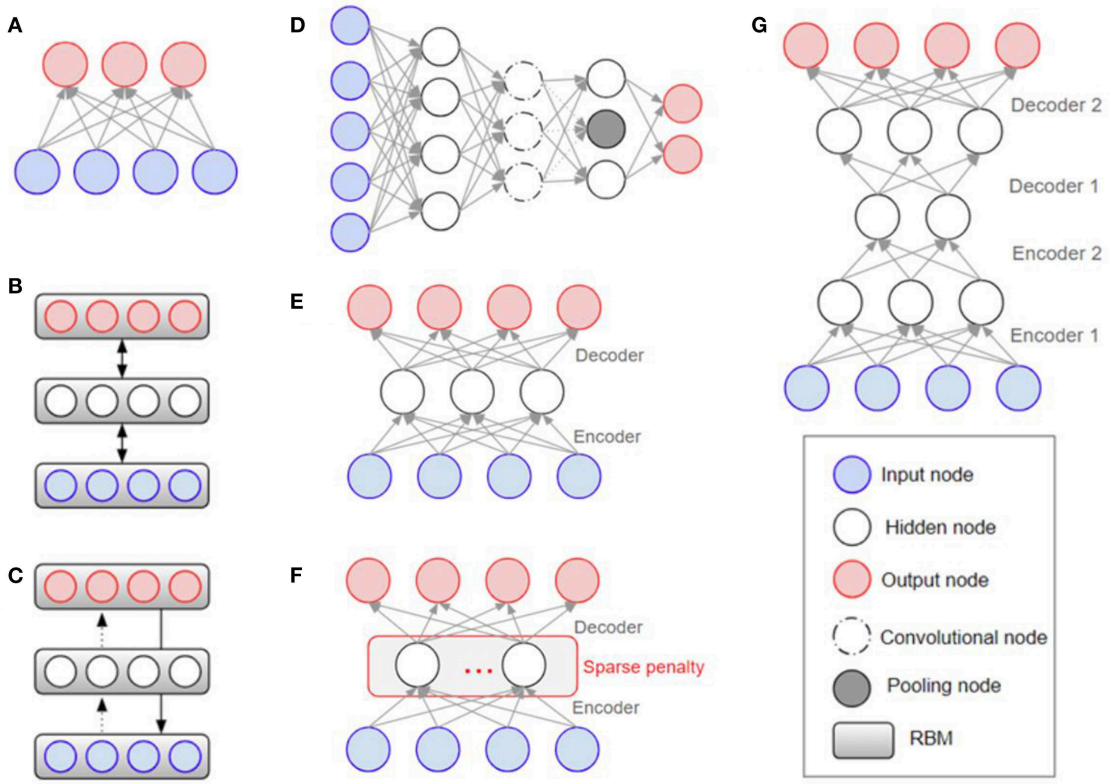

FIGURE 3 | Architectural structures in deep learning: **(A)** RBM (Hinton and Salakhutdinov, 2006) **(B)** DBM (Salakhutdinov and Larochelle, 2010) **(C)** DBN (Bengio, 2009) **(D)** CNN (Krizhevsky et al., 2012) **(E)** AE (Fukushima, 1975; Krizhevsky and Hinton, 2011) **(F)** Sparse AE (Vincent et al., 2008, 2010) **(G)** Stacked AE (Larochelle et al., 2007; Makhzani and Frey, 2015). RBM, Restricted Boltzmann Machine; DBM, Deep Boltzmann Machine; DBN, Deep Belief Network; CNN, Convolutional Neural Network; AE, Auto-Encoders.

TABLE 1 | Definition of acronyms.

Acronym	Description	Acronym	Description
ANN	Artificial neural network	CNN	Convolutional neural network
DNN	Deep neural network	RNN	Recurrent neural network
RBM	Restricted Boltzmann machine	GAN	Generative adversarial networks
DBM	Deep Boltzmann machine	SGD	Stochastic gradient descent
DBN	Deep belief network	SVM	Support vector machine
AE	Auto-encoders	ROI	Regions of interest
SAE	Stacked auto-encoder	HMM	Hidden markov model

the full text was not accessible. The remaining 192 articles were downloaded for review.

Eligibility Screening

Out of the 192 publications retrieved, 150 articles were excluded because the authors only introduced or mentioned deep learning but did not use it. Out of the 42 remaining publications, (1) 18 articles were excluded because they did not perform deep learning approaches for AD classification and/or prediction of MCI to AD conversion; (2) 5 articles were excluded because their neuroimaging data were not explicitly described; and (3) 3 articles were excluded because performance results were not provided. The remaining 16 papers were included in this review for AD classification and/or prediction of MCI to AD conversion. All of the final selected and compared papers used ADNI data in common.

RESULTS

From the 16 papers included in this review, **Table 2** provides the top results of diagnostic classification and/or prediction of MCI to AD conversion. We compared only binary classification results. Accuracy is a measure used consistently in the 16 publications. However, it is only one metric of the performance characteristics of an algorithm. The group composition, sample sizes, and number of scans analyzed are also noted together because accuracy is sensitive to unbalanced distributions. **Table S1** shows the full results sorted according to the performance accuracy as well as the number of subjects, the deep learning approach, and the neuroimaging type used in each paper.

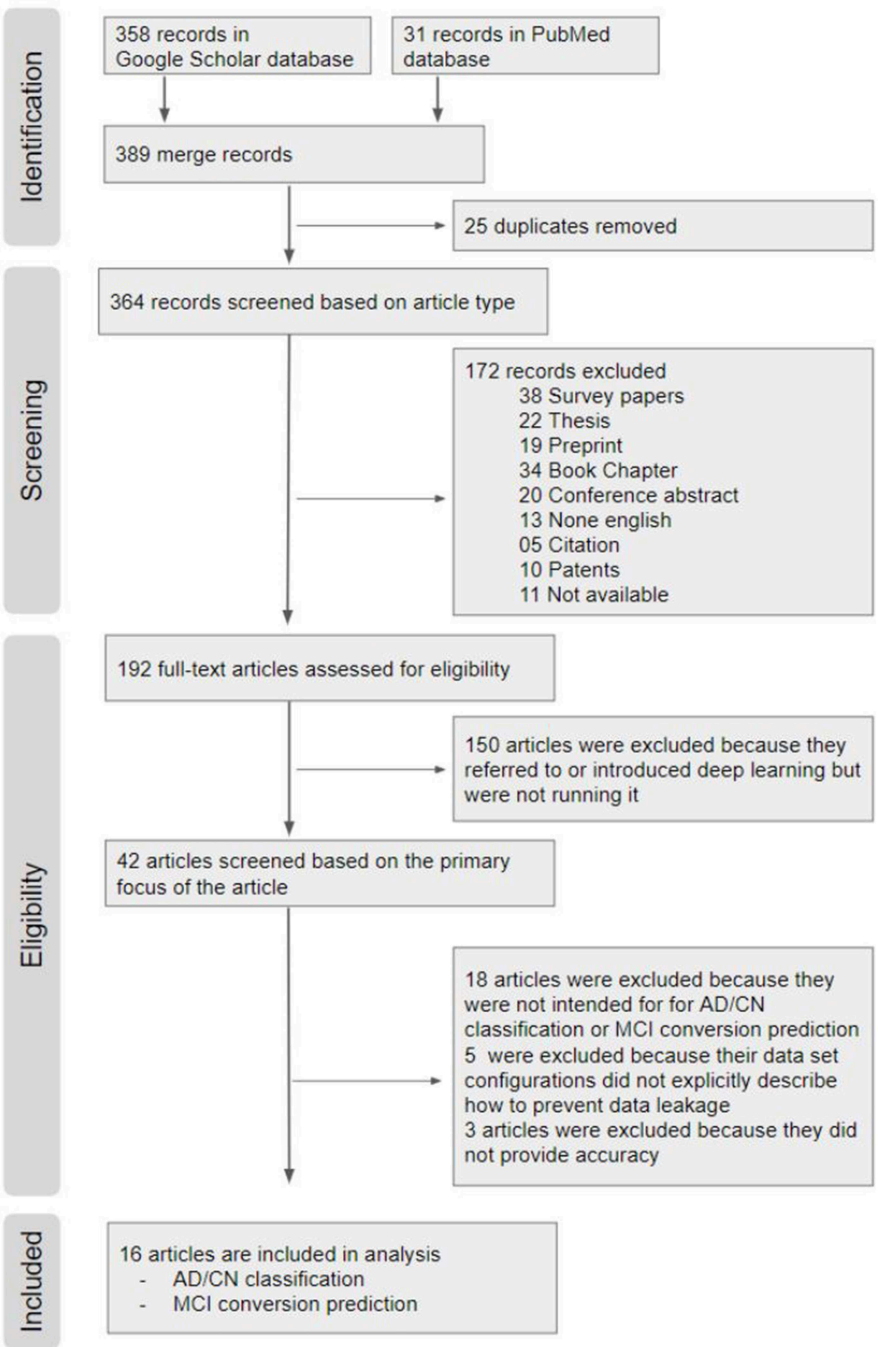

FIGURE 4 | PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Flow Chart. From a total of 389 hits on Google scholar and PubMed search, 16 articles were included in the systematic review.

Deep Learning for Feature Selection From Neuroimaging Data

Multimodal neuroimaging data have been used to identify structural and molecular/functional biomarkers for AD. It has been shown that volumes or cortical thicknesses in pre-selected AD-specific regions, such as the hippocampus and entorhinal cortex, could be used as features to enhance the classification accuracy in machine learning. Deep learning approaches have been used to select features from neuroimaging data.

As shown in **Figure 5**, 4 studies have used hybrid methods that combine deep learning for feature selection from neuroimaging data and traditional machine learning, such as the SVM as a

TABLE 2 | Summary of 16 previous studies to systematically be reviewed.

References	Modality	Data processing/training	Classifier	AD:NC acc.	SEN	SPE	cMCI:ncMCI acc.	SEN	SPE	AD	cMCI	ncMCI	NC	Total
Suk and Shen (2013)	MRI, PET, CSF	SAE	SVM	**95.9**			**75.8**			51	43	56	52	202
Liu et al. (2014)	MRI, PET	SAE + NN	Softmax	**87.76**	88.57	87.22	76.92 (MCI:NC)	74.29	78.13	65	67	102	77	311
Suk et al. (2014)	MRI, PET	DBM	SVM	**95.35**	94.65	95.22	**75.92** 86.75 (MCI:NC)	48.04 95.37	95.23 65.87	93	76	128	101	398
Li et al. (2014)	MRI, PET	3D CNN	Logistic regression	**92.87**			76.21 (MCI:NC)			198	167	236	229	830
Li et al. (2015)	MRI, PET, CSF	RBM + drop out	SVM	**91.4**			**57.4** 76.21 (MCI:NC)			51	43	56	52	202
Suk et al. (2015)	MRI, PET, CSF	SAE + sparse learning	SVM	**98.8**			**83.3** 90.7 (MCI:NC)			51	43	56	52	202
Liu et al. (2015)	MRI, PET	SAE with zero-masking	Softmax	**91.4**	92.32	90.42	82.1 (MCI:NC)	60.0	92.32	77	67	102	85	331
Cheng et al. (2017)	MRI	3D CNN	Softmax	**87.15**	86.36	85.93				199			229	428
Cheng and Liu (2017)	MRI, PET	3D CNN + 2D CNN	Softmax	**89.64**	87.10	92.00				93			100	193
Aderghal et al. (2017)	MRI	2D CNN	Softmax	**91.41**	93.75	89.06	65.62 (MCI:NC)	66.25	65.0	188	399 (MCI)		228	815
Korolev et al. (2017)	MRI	3D CNN	Softmax	**80**	87 (AUC)		61 (lMCI:NC) 56 (lMCI:NC)	65 (AUC) 58 (AUC)		50	43 (lMCI)	77 (eMCI)	61	111
Vu et al. (2017)	MRI, PET	SAE + 3D CNN	Softmax	**91.14**						145			172	317
Liu et al. (2018a)	PET	RNN	Softmax	**91.2**	91.4	91.0	78.9 (MCI:NC)	78.01	80.0	93	146 (MCI)		100	339
Liu et al. (2018b)	MRI	Landmark detection + 3D CNN	Softmax	**91.09**	88.05	93.50	**76.9**	42.11	82.43	159	38	239	200	636
Lu et al. (2018)	MRI, PET	DNN + NN	Softmax	**84.6**	80.2	91.8	**82.93**	79.69	83.84	238	217	409	360	1224
Choi and Jin (2018)	PET	3D CNN	Softmax	**96**	93.5	97.8	**84.2**	81.0	87.0	139	79	92	182	492

SEN = TP/(TP + FN), SPE = TN/(TN + FP). TP, true positive; TN, true negative; FP, false positive; FN, false negative. All data on this table were from ADNI.

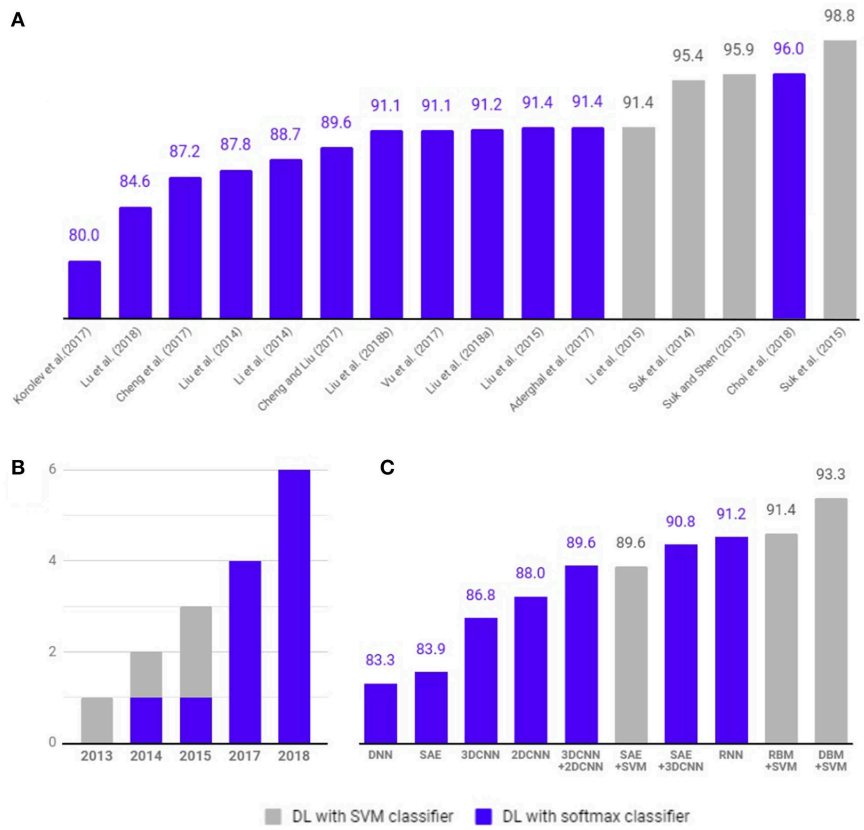

FIGURE 5 | Comparison of diagnostic classification accuracy of pure deep learning and hybrid approach. Four studies (gray) have used hybrid methods that combine deep learning for feature selection from neuroimaging data and traditional machine learning, such as the SVM as a classifier. Twelve studies (blue) have used deep learning method with softmax classifier for diagnostic classification and/or prediction of MCI to AD conversion. **(A)** Accuracy comparison between articles. **(B)** Number of studies published per year. **(C)** Average classification accuracy of each methods.

classifier. Suk and Shen (2013) used a stacked auto-encoder (SAE) to construct an augmented feature vector by concatenating the original features with outputs of the top hidden layer of the representative SAEs. Then, they used a multi-kernel SVM for classification to show 95.9% accuracy for AD/CN classification and 75.8% prediction accuracy of MCI to AD conversion. These methods successfully tuned the input data for the SVM classifier. However, SAE as a classifier (Suk et al., 2015) yielded 89.9% accuracy for AD/CN classification and 60.2% accuracy for prediction of MCI to AD conversion. Later Suk et al. (2015) extended the work to develop a two-step learning scheme: greedy layer-wise pre-training and fine-tuning in deep learning. The same authors further extended their work to use the DBM to find latent hierarchical feature representations by combining heterogeneous modalities during the feature representation learning (Suk et al., 2014). They obtained 95.35% accuracy for AD/CN classification and 74.58% prediction accuracy of MCI to AD conversion. In addition, the authors initialized SAE parameters with target-unrelated samples and tuned the optimal parameters with target-related samples to have 98.8% accuracy for AD/CN classification and 83.7% accuracy for prediction of MCI to AD conversion (Suk et al., 2015). Li et al. (2015) used the RBM with a dropout technique to reduce overfitting in deep learning and SVM as a classifier, which produced 91.4% accuracy for AD/CN classification and 57.4% prediction accuracy of MCI to AD conversion.

Deep Learning for Diagnostic Classification and Prognostic Prediction

To select optimal features from multimodal neuroimaging data for diagnostic classification, we usually need several pre-processing steps, such as neuroimaging registration and feature extraction, which greatly affect the classification performance. However, deep learning approaches have been applied to AD diagnostic classification using original neuroimaging data without any feature selection procedures.

As shown in **Figure 5**, 12 studies have used only deep learning for diagnostic classification and/or prediction of MCI to AD conversion. Liu et al. (2014) used stacked sparse auto-encoders (SAEs) and a softmax regression layer and showed 87.8% accuracy for AD/CN classification. Liu et al. (2015) used SAE and a softmax logistic regressor as well as a zero-mask strategy for data fusion to extract complementary information from multimodal neuroimaging data (Ngiam et al., 2011), where

one of the modalities is randomly hidden by replacing the input values with zero to converge different types of image data for SAE. Here, the deep learning algorithm improved accuracy for AD/CN classification by 91.4%. Recently, Lu et al. (2018) used SAE for pre-training and DNN in the last step, which achieved an AD/CN classification accuracy of 84.6% and an MCI conversion prediction accuracy of 82.93%. CNN, which has shown remarkable performance in the field of image recognition, has also been used for the diagnostic classification of AD with multimodal neuroimaging data. Cheng et al. (2017) used image patches to transform the local images into high-level features from the original MRI images for the 3D-CNN and yielded 87.2% accuracy for AD/CN classification. They improved the accuracy to 89.6% by running two 3D-CNNs on neuroimage patches extracted from MRI and PET separately and by combining their results to run 2D CNN (Cheng and Liu, 2017). Korolev et al. (2017) applied two different 3D CNN approaches [plain (VoxCNN) and residual neural networks (ResNet)] and reported 80% accuracy for AD/CN classification, which was the first study that the manual feature extraction step was unnecessary. Aderghal et al. (2017) captured 2D slices from the hippocampal region in the axial, sagittal, and coronal directions and applied 2D CNN to show 85.9% accuracy for AD/CN classification. Liu et al. (2018b) selected discriminative patches from MR images based on AD-related anatomical landmarks identified by a data-driven learning approach and ran 3D CNN on them. This approach used three independent data sets (ADNI-1 as training, ADNI-2 and MIRIAD as testing) to yield relatively high accuracies of 91.09 and 92.75% for AD/CN classification from ADNI-2 and MIRIAD, respectively, and an MCI conversion prediction accuracy of 76.9% from ADNI-2. Li et al. (2014) trained 3D CNN models on subjects with both MRI and PET scans to encode the non-linear relationship between MRI and PET images and then used the trained network to estimate the PET patterns for subjects with only MRI data. This study obtained an AD/CN classification accuracy of 92.87% and an MCI conversion prediction accuracy of 72.44%. Vu et al. (2017) applied SAE and 3D CNN to subjects with MRI and FDG PET scans to yield an AD/CN classification accuracy of 91.1%. Liu et al. (2018a) decomposed 3D PET images into a sequence of 2D slices and used a combination of 2D CNN and RNNs to learn the intra-slice and inter-slice features for classification, respectively. The approach yielded AD/CN classification accuracy of 91.2%. If the data is imbalanced, the chance of misdiagnosis increases and sensitivity decreases. For example, in Suk et al. (2014) there were 76 cMCI and 128 ncMCI subjects and the obtained sensitivity of 48.04% was low. Similarly, Liu et al. (2018b) included 38 cMCI and 239 ncMCI subjects and had a low sensitivity of 42.11%. Recently Choi and Jin (2018) reported the first use of 3D CNN models to multimodal PET images [FDG PET and [18F]florbetapir PET] and obtained 96.0% accuracy for AD/CN classification and 84.2% accuracy for the prediction of MCI to AD conversion.

Performance Comparison by Types of Neuroimaging Techniques

In order to improve the performance for AD/CN classification and for the prediction of MCI to AD conversion, multimodal neuroimaging data such as MRI and PET have commonly been used in deep learning: MRI for brain structural atrophy, amyloid PET for brain amyloid-β accumulation, and FDG-PET for brain glucose metabolism. MRI scans were used in 13 studies, FDG-PET scans in 10, both MRI and FDG-PET scans in 12, and both amyloid PET and FDG-PET scans in 1. The performance in AD/CN classification and/or prediction of MCI to AD conversion yielded better results in PET data compared to MRI. Two or more multimodal neuroimaging data types produced

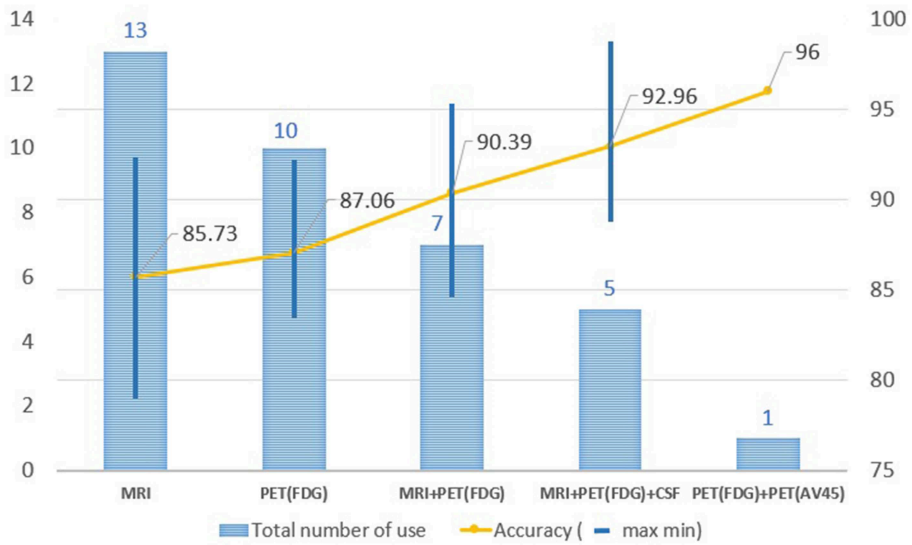

FIGURE 6 | Changes in accuracy by types of image resource. MRI scans were used in 13 studies, FDG-PET scans in 10, both MRI and FDG-PET scans in 12, and both amyloid PET and FDG-PET scans in 1. The performance in AD/CN classification yielded better results in PET data compared to MRI. Two or more multimodal neuroimaging data types produced higher accuracies than a single neuroimaging technique.

higher accuracies than a single neuroimaging technique. **Figure 6** shows the results of the performance comparison by types of neuroimaging techniques.

Performance Comparison by Deep Learning Algorithms

Deep learning approaches require massive amounts of data to achieve the desired levels of performance accuracy. In currently limited neuroimaging data, the hybrid methods that combine traditional machine learning methods for diagnostic classification with deep learning approaches for feature extraction yielded better performance and can be a good alternative to handle the limited data. Here, an auto-encoder (AE) was used to decode the original image values, making them similar to the original image, which it then included as input, thereby effectively utilizing the limited neuroimaging data. Although hybrid approaches have yielded relatively good results, they do not take full advantage of deep learning, which automatically extracts features from large amounts of neuroimaging data. The most commonly used deep learning method in computer vision studies is the CNN, which specializes in extracting characteristics from images. Recently, 3D CNN models using multimodal PET images [FDG-PET and [18F]florbetapir PET] showed better performance for AD/CN classification and for the prediction of MCI to AD conversion.

DISCUSSION

Effective and accurate diagnosis of Alzheimer's disease (AD) is important for initiation of effective treatment. Particularly, early diagnosis of AD plays a significant role in therapeutic development and ultimately for effective patient care. In this study, we performed a systematic review of deep learning approaches based on neuroimaging data for diagnostic classification of AD. We analyzed 16 articles published between 2013 and 2018 and classified them according to deep learning algorithms and neuroimaging types. Among 16 papers, 4 studies used a hybrid method to combine deep learning and traditional machine learning approaches as a classifier, and 12 studies used only deep learning approaches. In a limited available neuroimaging data set, hybrid methods have produced accuracies of up to 98.8% for AD classification and 83.7% for prediction of conversion from MCI to AD. Deep learning approaches have yielded accuracies of up to 96.0% for AD classification and 84.2% for MCI conversion prediction. While it is a source of concern when experiments obtain a high accuracy using small amounts of data, especially if the method is vulnerable to overfitting, the highest accuracy of 98.8% was due to the SAE procedure, whereas the 96% accuracy was due to the amyloid PET scan, which included pathophysiological information regarding AD. The highest accuracy for the AD classification was 87% when 3DCNN was applied from the MRI without the feature extraction step (Cheng et al., 2017). Therefore, two or more multimodal neuroimaging data types have been shown to produce higher accuracies than a single neuroimaging type.

In traditional machine learning, well-defined features influence performance results. However, the greater the complexity of the data, the more difficult it is to select optimal features. Deep learning identifies optimal features automatically from the data (i.e., the classifier trained by deep learning finds features that have an impact on diagnostic classification without human intervention). Because of its ease-of-use and better performance, deep learning has been used increasingly for medical image analysis. The number of studies of AD using CNN, which show better performance in image recognition among deep learning algorithms, has increased drastically since 2015. This is consistent with a previous survey showing that the use of deep learning for lesion classification, detection, and segmentation has also increased rapidly since 2015 (Litjens et al., 2017).

Recent trends in the use of deep learning are aimed at faster analysis with better accuracy than human practitioners. Google's well-known study for the diagnostic classification of diabetic retinopathy (Gulshan et al., 2016) showed classification performance that goes well beyond that of a skilled professional. The diagnostic classification by deep learning needs to show consistent performance under various conditions, and the predicted classifier should be interpretable. In order for diagnostic classification and prognostic prediction using deep learning to reach readiness for real world clinical applicability, several issues need to be addressed, as discuss below.

Transparency

Traditional machine learning approaches may require expert involvement in preprocessing steps for feature extraction and selection from images. However, since deep learning does not require human intervention but instead extracts features directly from the input images, the data preprocessing procedure is not routinely necessary, allowing flexibility in the extraction of properties based on various data-driven inputs. Therefore, deep learning can create a good, qualified model at each time of the run. The flexibility has shown deep learning to achieve a better performance than other traditional machine learning that relies on preprocessing (Bengio, 2013). However, this aspect of deep learning necessarily brings uncertainty over which features would be extracted at every epoch, and unless there is a special design for the feature, it is very difficult to show which specific features were extracted within the networks (Goodfellow et al., 2016). Due to the complexity of the deep learning algorithm, which has multiple hidden layers, it is also difficult to determine how those selected features lead to a conclusion and to the relative importance of specific features or subclasses of features. This is a major limitation for mechanistic studies where understanding the informativeness of specific features is desirable for model building. These uncertainties and complexities tend to make the process of achieving high accuracy opaque and also make it more difficult to correct any biases that arise from a given data set. This lack of clarity also limits the applicability of obtained results to other use cases.

The issue of transparency is linked to the clarity of the results from machine learning and is not a problem limited to deep learning (Kononenko, 2001). Despite the simple principle, the complexity of the algorithm makes it difficult to describe mathematically. When one perceptron advances to a neural network by adding more hidden layers, it becomes even more difficult to explain why a particular prediction was made. AD classification based on 3D multimodal medical images with deep learning involves non-linear convolutional layers and pooling that have different dimensionality from the source data, making it very difficult to interpret the relative importance of discriminating features in original data space. This is a fundamental challenge in view of the importance of anatomy in the interpretation of medical images, such as MRI or PET scans. The more advanced algorithm generates plausible results, but the mathematical background is difficult to explain, although the output for diagnostic classification should be clear and understandable.

Reproducibility

Deep learning performance is sensitive to the random numbers generated at the start of training, and hyper-parameters, such as learning rates, batch sizes, weight decay, momentum, and dropout probabilities, may be tuned by practitioners (Hutson, 2018). To produce the same experimental result, it is important to set the same random seeds on multiple levels. It is also important to maintain the same code bases (Vaswani et al., 2018), even though the hyper-parameters and random seeds were not, in most cases, provided in our study. The uncertainty of the configuration and the randomness involved in the training procedure may make it difficult to reproduce the study and achieve the same results.

When the available neuroimaging data is limited, careful consideration at the architectural level is needed to avoid the issues of overfitting and reproducibility. Data leakage in machine learning (Smialowski et al., 2009) occurs when the data set framework is designed incorrectly, resulting in a model that uses inessential additional information for classification. In the case of diagnostic classification for the progressive and irreversible Alzheimer's disease, all subsequent MRI images should be labeled as belonging to a patient with Alzheimer's disease. Once the brain structure of the patient is shared by both the training and testing sets, the morphological features of the patient's brain greatly influence the classification decision, rather than the biomarkers of dementia. In the present study, articles were excluded from the review if the data set configurations did not explicitly describe how to prevent data leakage (**Figure 4**).

Future studies ultimately need to replicate key findings from deep learning on entirely independent data sets. This is now widely recognized in genetics (König, 2011; Bush and Moore, 2012) and other fields but has been slow to penetrate deep learning studies employing neuroimaging data. Hopefully the emerging open ecology of medical research data, especially in the AD and related disorders field (Toga et al., 2016; Reas, 2018), will provide a basis to remediate this problem.

OUTLOOK AND FUTURE DIRECTION

Deep Learning algorithms and applications continue to evolve, producing the best performance in closed-ended cases, such as image recognition (Marcus, 2018). It works particularly well when inference is valid, i.e., the training and test environments are similar. This is especially true in the study of AD when using neuroimages (Litjens et al., 2017). One weakness of deep learning is that it is difficult to modify potential bias in the network when the complexity is too great to guarantee transparency and reproducibility. The issue may be solved through the accumulation of large-scale neuroimaging data and by studying the relationships between deep learning and features. Disclosing the parameters used to obtain the results and mean values from sufficient experimentations can mitigate the issue of reproducibility.

Not all problems can be solved with deep learning. Deep learning that extracts attributes directly from the input data without preprocessing for feature selection has difficulty integrating different formats of data as an input, such as neuroimaging and genetic data. Because the adjustment of weights for the input data is performed automatically within a closed network, adding additional input data into the closed network causes confusion and ambiguity. A hybrid approach, however, puts the additional information into machine learning parts and the neuroimages into deep learning parts before combining the two results.

Progress will be made in deep learning by overcoming these issues while presenting problem-specific solutions. As more and more data are acquired, research using deep learning will become more impactful. The expansion of 2D CNN into 3D CNN is important, especially in the study of AD, which deals with multimodal neuroimages. In addition, Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) may be applicable for generating synthetic medical images for data augmentation. Furthermore, reinforcement learning (Sutton and Barto, 2018), a form of learning that adapts to changes in data as it makes its own decision based on the environment, may also demonstrate applicability in the field of medicine.

AD research using deep learning is still evolving to achieve better performance and transparency. As multimodal neuroimaging data and computer resources grow rapidly, research on the diagnostic classification of AD using deep learning is shifting toward a model that uses only deep learning algorithms rather than hybrid methods, although methods need to be developed to integrate completely different formats of data in a deep learning network.

DATA AVAILABILITY

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

TJ and AS: conceptualization and study design. TJ: data collection and analysis and drafting manuscript. TJ, KN, and AS: revision of the manuscript for important scientific content and final approval.

FUNDING

This review was supported, in part, by grants from the National Institutes of Health (NIH) and include the following sources: P30 AG10133, R01 AG19771, R01 AG057739, R01 CA129769, R01 LM012535, and R03 AG054936. Many studies reviewed here analyzed data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) that was funded by the National Institutes of Health (U01 AG024904) and Department of Defense (W81XWH-12-2-0012) and a consortium of private partners.

ACKNOWLEDGMENTS

We are grateful to all of the study participants and their families that participated in the neuroimaging research on Alzheimer's disease reviewed here. We are also indebted to the clinical and computational researchers who reported their results, facilitating the analyses and discussion in this systematic review. We thank Paula J. Bice, Ph.D., for editorial assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnagi.2019.00220/full#supplementary-material

REFERENCES

Aderghal, K., Benois-Pineau, J., Afdel, K., and Catheline, G. (2017). "FuseMe: classification of sMRI images by fusion of deep CNNs in 2D+ϵ projections," in *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing* (New York, NY).

Alzheimer's Association (2018). 2018 Alzheimer's disease facts and figures. *Alzheimer's Dementia* 14, 367–429. doi: 10.1016/j.jalz.2018.02.001

Bengio, Y. (2009). Learning deep architectures for AI. *Found. Trends Mach. Learn.* 2, 1–127. doi: 10.1561/2200000006

Bengio, Y. (2013). "Deep learning of representations: looking forward," in *International Conference on Statistical Language and Speech Processing* (Tarragona: Springer), 1–37.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* New York, NY: Oxford University Press.

Bottou, L. (2010). "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010* (Paris: Springer), 177–186.

Boureau, Y.-L., Ponce, J., and Lecun, Y. (2010). "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Haifa), 111–118.

Bush, W. S., and Moore, J. H. (2012). Genome-wide association studies. *PLoS Comput. Biol.* 8:e1002822. doi: 10.1371/journal.pcbi.1002822

Cheng, D., and Liu, M. (2017). "CNNs based multi-modality classification for AD diagnosis," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)* (Shanghai), 1–5.

Cheng, D., Liu, M., Fu, J., and Wang, Y. (2017). "Classification of MR brain images by combination of multi-CNNs for AD diagnosis," in *Ninth International Conference on Digital Image Processing (ICDIP 2017)* (Hong Kong: SPIE), 5.

Choi, H., and Jin, K. H. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behav. Brain Res.* 344, 103–109. doi: 10.1016/j.bbr.2018.02.017

Ciregan, D., Meier, U., and Schmidhuber, J. (2012). "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI), 3642–3649.

De strooper, B., and Karran, E. (2016). The cellular phase of Alzheimer's disease. *Cell* 164, 603–615. doi: 10.1016/j.cell.2015.12.056

Farabet, C., Couprie, C., Najman, L., and Lecun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1915–1929. doi: 10.1109/TPAMI.2012.231

Fukushima, K. (1975). Cognitron: a self-organizing multilayered neural network. *Biol. Cybernet.* 20, 121–136. doi: 10.1007/BF00342633

Fukushima, K. (1979). Neural network model for a mechanism of pattern recognition unaffected by shift in position-Neocognitron. *IEICE Tech. Rep. A* 62, 658–665.

Fukushima, K. (1980). Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202. doi: 10.1007/BF00344251

Galvin, J. E. (2017). Prevention of Alzheimer's disease: lessons learned and applied. *J. Am. Geriatr. Soc.* 65, 2128–2133. doi: 10.1111/jgs.14997

Glorot, X., Bordes, A., and Bengio, Y. (2011). "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics* (Fort Lauderdale, FL), 315–323.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning.* Cambridge, MA: MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27,* eds Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Montreal, QC), 2672–2680.

Gulshan, V., Peng, L., Coram, M., Stumpe MC, Wu D, Narayanaswamy A, et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 2402–2410. doi: 10.1001/jama.2016.17216

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-R., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.* 29, 82–97. doi: 10.1109/MSP.2012.2205597

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* 18, 1527–1554. doi: 10.1162/neco.2006.18.7.1527

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hinton, G. E., and Zemel, R. S. (1994). "Autoencoders, minimum description length and Helmholtz free energy," in *Advances in Neural Information Processing Systems 6*, eds J. D. Cowan, G. Tesauro, and J. Alspector (Denver, CO), 3–10.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science* 359, 725–726. doi: 10.1126/science.359.6377.725

Ivakhnenko, A. G. (1968). The group method of data of handling; a rival of the method of stochastic approximation. *Sov. Autom. Control* 13, 43–55.

Ivakhnenko, A. G. (1971). Polynomial theory of complex systems. *IEEE Trans. Syst. Man Cybern.* SMC-1, 364–378. doi: 10.1109/TSMC.1971.4308320

Ivakhnenko, A. G. E., and Lapa, V. G. (1965). *Cybernetic Predicting Devices.* New York, NY: CCM Information Corporation.

König, I. R. (2011). Validation in genetic association studies. *Brief. Bioinformatics* 12, 253–258. doi: 10.1093/bib/bbq074

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.* 23, 89–109. doi: 10.1016/S0933-3657(01)00077-X

Korolev, S., Safiullin, A., Belyaev, M., and Dodonova, Y. (2017). "Residual and plain convolutional neural networks for 3D brain MRI classification," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (Melbourne, VIC), 835–838.

Krizhevsky, A., and Hinton, G. E. (2011). "Using very deep autoencoders for content-based image retrieval," in *Proceedings of the 19th European Symposium on Artificial Neural Networks: ESANN 2011* (Bruges), 2.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Stateline, NV), 1097–1105.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007). "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the 24th International Conference on Machine Learning* (Corvallis, OR: ACM), 473–480.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521:436. doi: 10.1038/nature14539

Lecun, Y., Touresky, D., Hinton, G., and Sejnowski, T. (1988). "A theoretical framework for back-propagation," in *Proceedings of the 1988 Connectionist Models Summer School: CMU* (Pittsburgh, PA: Morgan Kaufmann), 21–28.

Li, F., Tran, L., Thung, K.-H., Ji, S., Shen, D., and Li, J. (2015). A robust deep model for improved classification of AD/MCI patients. *IEEE J. Biomed. Health Inform.* 19, 1610–1616. doi: 10.1109/JBHI.2015.2429556

Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D., et al. (2014). "Deep learning based imaging data completion for improved brain disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol 17* (Boston, MA), 305–312.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, M., Cheng, D., Yan, W., and Alzheimer's Disease Neuroimaging Initiative. (2018a). Classification of Alzheimer's disease by combination of convolutional and recurrent neural networks using FDG-PET images. *Front. Neuroinform.* 12:35. doi: 10.3389/fninf.2018.00035

Liu, M., Zhang, J., Adeli, E., and Shen, D. (2018b). Landmark-based deep multi-instance learning for brain disease diagnosis. *Med. Image Anal.* 43, 157–168. doi: 10.1016/j.media.2017.10.005

Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., et al. (2015). Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's Disease. *IEEE Trans. Biomed. Eng.* 62, 1132–1140. doi: 10.1109/TBME.2014.2372011

Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., and Feng, D. (2014). "Early diagnosis of Alzheimer's disease with deep learning," in *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (Beijing), 1015–1018.

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., and Beg, M. F. (2018). Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. *Sci. Rep.* 8:5697. doi: 10.1038/s41598-018-22871-z

Lu, D., and Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *Int. J. Remote Sens.* 28, 823–870. doi: 10.1080/01431160600746456

Makhzani, A., and Frey, B. (2015). "k-sparse autoencoders," in *Advances in Neural Information Processing Systems 28* (Montreal, QC), 2791–2799.

Marcus, G. (2018). Deep learning: a critical appraisal. *arXiv preprint*. arXiv:1801.00631.

Mcculloch, W. S., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133. doi: 10.1007/BF02478259

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26,* eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Stateline, NV), 3111–3119.

Minsky, M., and Papert, S. (1969). *Perceptrons*. Cambridge, MA: MIT Press.

Moher, D., Liberati, A., Tetzlaff, J., and Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Ann. Intern. Med.* 151, 264–269. doi: 10.7326/0003-4819-151-4-200908180-00135

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Haifa), 807–814.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (Bellevue), 689–696.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., and Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage* 155, 530–548. doi: 10.1016/j.neuroimage.2017.03.057

Reas, E. (2018). ADNI: understanding Alzheimer's disease through collaboration and data sharing. *PLoS Blogs*. Retrieved from: https://blogs.plos.org/neuro/2018/10/24/adni-understanding-alzheimers-disease-through-collaboration-and-data-sharing/ (accessed October 24, 2018).

Riedel, B. C., Daianu, M., Ver Steeg, G., Mezher, A., Salminen, L. E., Galstyan, A., et al. (2018). Uncovering biologically coherent peripheral signatures of health and risk for Alzheimer's disease in the aging brain. *Front. Aging Neurosci.* 10:390. doi: 10.3389/fnagi.2018.00390

Ripley, B. D., and Hjort, N. (1996). *Pattern Recognition and Neural Networks*. New York, NY: Cambridge University Press.

Rosenblatt, F. (1957). *The Perceptron, A Perceiving and Recognizing Automaton Project Para.* Buffalo, NY: Cornell Aeronautical Laboratory.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65:386. doi: 10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323:533. doi: 10.1038/323533a0

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comp. Vision* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Salakhutdinov, R., and Larochelle, H. (2010). "Efficient learning of deep Boltzmann machines," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Sardinia), 693–700.

Samper-Gonzalez, J., Burgos, N., Bottani, S., Fontanella, S., Lu, P., Marcoux, A., et al. (2018). Reproducible evaluation of classification methods in Alzheimer's disease: framework and application to MRI and PET data. *Neuroimage* 183, 504–521. doi: 10.1016/j.neuroimage.2018.08.042

Schalkoff, R. J. (1997). *Artificial Neural Networks.* New York, NY:McGraw-Hill.

Schelke, M. W., Attia, P., Palenchar, D. J., Kaplan, B., Mureb, M., Ganzer, C. A., et al. (2018). Mechanisms of risk reduction in the clinical practice of Alzheimer's disease prevention. *Front. Aging Neurosci.* 10:96. doi: 10.3389/fnagi.2018.00096

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Smialowski, P., Frishman, D., and Kramer, S. (2009). Pitfalls of supervised feature selection. *Bioinformatics* 26, 440–443. doi: 10.1093/bioinformatics/btp621

Suk, H.-I., Lee, S.-W., Shen, D., and Alzheimer's Disease Neuroimaging Initiative. (2014). Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage* 101, 569–582. doi: 10.1016/j.neuroimage.2014.06.077

Suk, H.-I., Lee, S.-W., Shen, D., and The Alzheimer's Disease Neuroimaging, I. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct. Funct.* 220, 841–859. doi: 10.1007/s00429-013-0687-3

Suk, H.-I., and Shen, D. (2013). "Deep learning-based feature representation for AD/MCI classification," in *International Conference on Medical Image Computing and Computer-Assisted Intervention, Vol 16* (Nagoya), 583–590.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). "On the importance of initialization and momentum in deep learning," in *International Conference on Machine Learning* (Atlanta), 1139–1147.

Sutton, R. S., and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press.

Toga, A. W., Bhatt, P., and Ashish, N. (2016). Global data sharing in Alzheimer's disease research. *Alzheimer Dis. Assoc. Disord.* 30:160. doi: 10.1097/WAD.0000000000000121

Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., et al. (2018). "Tensor2tensor for neural machine translation," in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas* (Boston, MA), 193–199.

Veitch, D. P., Weiner, M. W., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2019). Understanding disease progression and improving Alzheimer's disease clinical trials: recent highlights from the Alzheimer's disease neuroimaging initiative. *Alzheimers Dement* 15, 106–152. doi: 10.1016/j.jalz.2018.08.005

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. (2008). "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning* (Indianapolis, IN: ACM), 1096–1103.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P. A. (2010). Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* 11, 3371–3408.

Vu, T. D., Yang, H.-J., Nguyen, V. Q., Oh, A. R., and Kim, M.-S. (2017). "Multimodal learning using convolution neural network and Sparse Autoencoder," in *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)* (Jeju), 309–312.

Werbos, P. J. (1982). "Applications of advances in nonlinear sensitivity analysis," in *System Modeling and Optimization,* eds R. F. Drenick and F. Kozin (New York, NY: Springer), 762–770.

Werbos, P. J. (2006). "Backwards differentiation in AD and neural nets: past links and new opportunities," in *Automatic Differentiation: Applications, Theory, and Implementations,* eds H. M. Bücker, G. Corliss, P. Hovland, U. Naumann, and B. Norris (New York, NY: Springer), 15–34.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership