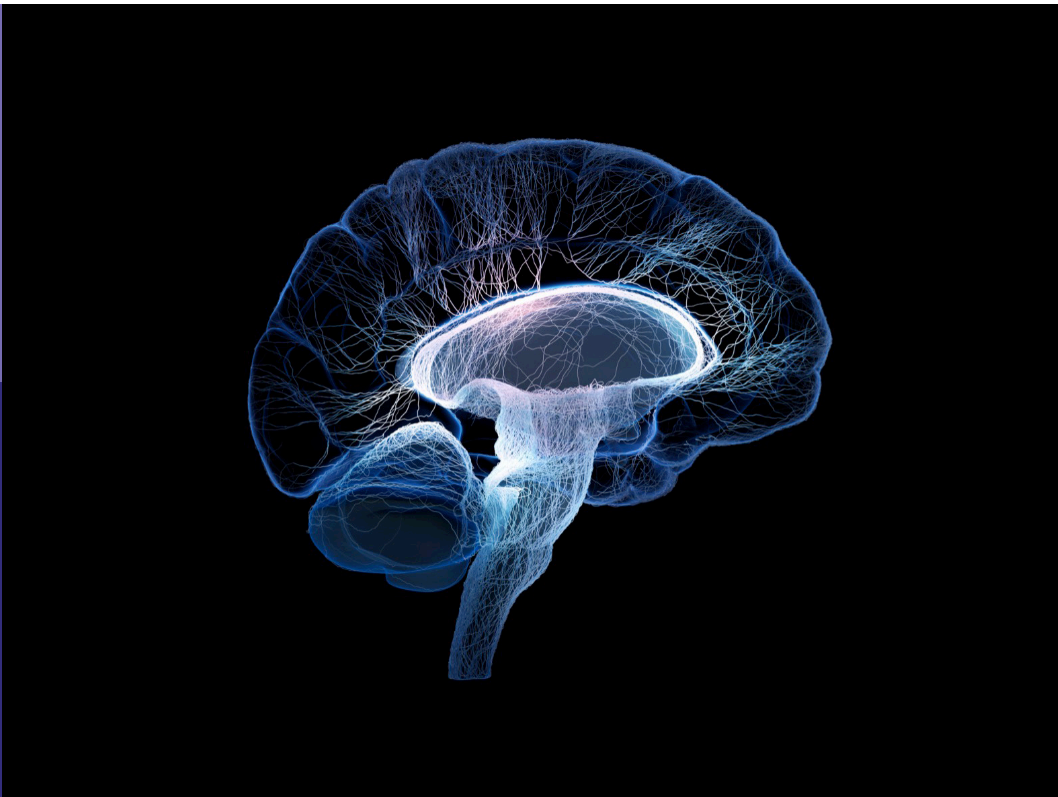
frontiers | Research Topics

Deep learning techniques and their applications to the healthy and disordered brain - during development through adulthood and beyond

Edited by
Amir Shmuel, Albert Yang, Yogesh Rathi and Hyunjin Park

Published in
Frontiers in Neuroscience



ISSN 1664-8714
ISBN 978-2-83251-380-4
DOI 10.3389/978-2-83251-380-4

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: frontiersin.org/about/contact

Deep learning techniques and their applications to the healthy and disordered brain - during development through adulthood and beyond

Topic editors

Amir Shmuel — McGill University, Canada
Albert Yang — National Yang Ming Chiao Tung University, Taiwan
Yogesh Rathi — Harvard Medical School, United States
Hyunjin Park — Sungkyunkwan University, Republic of Korea

Citation

Shmuel, A., Yang, A., Rathi, Y., Park, H., eds. (2023). *Deep learning techniques and their applications to the healthy and disordered brain - during development through adulthood and beyond*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-83251-380-4

Table of contents

frontiers | Frontiers in Neuroscience

TYPE Editorial
PUBLISHED 22 December 2022
DOI 10.3389/fnins.2022.1118233

OPEN ACCESS

EDITED AND REVIEWED BY
Vince D. Calhoun,
Georgia State University, United States

*CORRESPONDENCE
Amir Shmuel
✉ amir.shmuel@mcgill.ca

SPECIALTY SECTION
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

RECEIVED 07 December 2022
ACCEPTED 08 December 2022
PUBLISHED 22 December 2022

CITATION
Shmuel A, Park H, Rathi Y and Yang A (2022) Editorial: Deep learning techniques and their applications to the healthy and disordered brain - during development through adulthood and beyond. *Front. Neurosci.* 16:1118233. doi: 10.3389/fnins.2022.1118233

Editorial: Deep learning techniques and their applications to the healthy and disordered brain - during development through adulthood and beyond

Amir Shmuel[1,2,3]*, Hyunjin Park[4,5], Yogesh Rathi[6,7] and Albert Yang[8]

[1]Department of Neurology, McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, QC, Canada, [2]Department of Neurosurgery, McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, QC, Canada, [3]Department of Physiology and Biomedical Engineering, McConnell Brain Imaging Centre, Montreal Neurological Institute, McGill University, Montreal, QC, Canada, [4]Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, South Korea, [5]Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea, [6]Department of Psychiatry, Harvard Medical School, Boston, MA, United States, [7]Department of Radiology, Harvard Medical School, Boston, MA, United States, [8]Institute of Brain Science, National Yang-Ming Chiao Tung University, Taipei, Taiwan

KEYWORDS

deep learning, medical image segmentation, neuroimaging-based diagnosis, Alzheimer's disease, Parkinson's disease, multiple sclerosis, schizophrenia, autism

Editorial on the Research Topic
Deep learning techniques and their applications to the healthy and disordered brain - during development through adulthood and beyond

Data acquisition methods used in medical imaging have been developing at an unprecedented pace; however, the interpretation of the data and their use for identifying and detecting biomarkers of disease can be difficult. The use of standardized computational aids has been proven very effective in overcoming this difficulty. Deep learning (DL) is a sub-group of machine learning algorithms capable of automatically extracting discriminatory features from raw input. This capacity makes DL a powerful technique that is already transforming neuroimaging data acquisition, modeling, and analysis. DL is a key player in MRI image reconstruction of sub-sampled *k*-space (Zeng et al., 2021), making it possible to reduce the duration of data acquisition. DL is capable of domain-specific image processing, with example applications in the retrospective correction of movement artifacts (Küstner et al., 2019) and monitoring and quality control of large MRI databases (Pizarro et al., 2019). Importantly, DL has been showing promise in the early diagnosis of several diseases and disorders, including autism spectrum disorder, Alzheimer's disease, and Parkinson's disease (Feng et al., 2022). The potential of DL to be useful also to basic research—not necessarily related to diagnosis—is high. The aim of this Research Topic is to highlight the potential of applying DL techniques to neuroimaging methods and applications.

The applicability of deep learning to a wide range of neuroimaging data acquisition methods is reflected in the type of data presented in the Research Topic, including structural MRI, diffusion MRI, EEG, and microscopy data. Similarly, deep learning is applicable to a wide range of neurological, and neurodevelopmental conditions studied by neuroimaging. Indeed, the Research Topic includes studies that apply deep learning to data on Alzheimer's disease, vascular cognitive impairment, Parkinson's disease, multiple sclerosis, autism, and neurodevelopment in pre-term infants possibly leading to cognitive deficits. The only psychiatric condition studied in this Research Topic is schizophrenia. However, thanks to its capacity to detect bio-markers that were not pre-defined, we expect that deep learning will be central to advancing biomarker identification of all psychiatric conditions.

Deep learning neural networks are especially capable of performing analysis of structured data, such as images and volumes. The Research Topic includes five studies that developed and/or evaluated deep learning methods for image segmentation. Brusini et al. **(ms. 469755) seek ways to improve the segmentation of the hippocampus** in structural MRI images. The structural integrity and volume of the hippocampus have been implicated as a biomarker in neurodegenerative conditions including Alzheimer's disease. They propose a DL-based hippocampus segmentation framework that embeds the statistical shape of the hippocampus as context information for learning. Their results suggest that adding shape information can improve the segmentation accuracy in cross-cohort validation, i.e., when deep neural networks are trained on one cohort and applied to another. **Another brain region implicated in neurological and psychiatric conditions is the amygdala**, whose segmentation is challenging due to its small dimensions. Large image patches are more likely to be dominated by background voxels, creating a class imbalance between the background class and the class of the small amygdala. Segmenting small structures such as the amygdala introduces a trade-off between capturing a sufficiently large context and retaining fine details while alleviating the imbalanced class issue. Alexander et al. **(ms. 497969)** addressed this challenging task by developing a dual-branch dilated residual 3D fully convolutional network (i.e., a network that performs only convolution, sub-sampling, and up-sampling) using receptive fields at the approximate size of the regions of interest with parallel convolutions to extract global context. **Segmenting the neonatal cerebrum according to tissue type is challenging** given its uniquely inverted tissue contrasts. Existing neuroimaging analysis packages are primarily designed to work on MRI with adult contrast but inversed water-to-cholesterol ratio in newborns leads to inverted MRI tissue contrast, hindering analyses. Ding et al. **(ms. 493147)** evaluated the performance of two architectures on segmenting T1 and T2 MRI images of the neonatal brain according to tissue types. HyperDense-Net performed better than LiviaNET, although it required a longer duration of training. Hong et al. **(ms. 591683)** developed a DL architecture for **segmenting MRI images of the fetal cortical plate during development**. They propose a fully convolutional neural network with a novel hybrid loss function and multi-view (axial, coronal, and sagittal) aggregation using a test-time augmentation, enabling the use of three-dimensional (3D) information. They demonstrate that these methods improve the accuracy of cortical plate segmentation. Closing the section on segmentation, Tan et al. **(ms. 481187) introduce DeepBrainSeg, a convolutional neural network for segmenting optical microscope images**. The classical method for parcellating the brain and the cerebral cortex relies on microscopic differences in neurons' size, density, and cortical myelin content, observed through a microscope. Parcellation is essential for the analysis of brain structures and their functions. DeepBrainSeg incorporates three feature levels to learn local and contextual features in different receptive fields through a dual-pathway convolutional neural network. It has been applied to mouse brains but is likely to obtain similar results if applied to larger brains.

Machine learning and deep learning carry the potential of distinguishing healthy brains and brains with neurological or psychiatric conditions, and diagnosing the conditions. Zhang et al. **(ms. 560709)** ***present A*** **survey on deep learning for neuroimaging-based brain disorder analysis**. They provide an overview of deep learning techniques and popular network architectures, and deep learning methods for computer-aided analysis of Alzheimer's disease, Parkinson's disease, Autism spectrum disorder, and Schizophrenia. They also discuss the limitations of existing studies and present possible future directions. Yamaguchi et al. **(ms. 652987)** take on the challenge of overcoming one of the current limitations. They demonstrate that a **3D convolutional autoencoder applied to structural MRI images of schizophrenia patients can extract features related to schizophrenia without relying on diagnostic labels**. They demonstrate that the proposed auto-encoder extracted features retained information that could predict medication dose and symptom severity in schizophrenia. Feature extraction without using diagnostic labels based on the current diagnostic criteria may lead to the development of alternative data-driven diagnostic criteria and could have a significant contribution to neuroimaging of neurological and psychiatric conditions. Another neurological condition whose early diagnosis and classification into sub-types can inform a decision on treatment is **Subcortical Vascular Cognitive Impairment**. Chen Q. et al. **(ms. 543607)** propose a deep learning solution using 3D attention-based Resnet applied to single T2-weighted FLAIR MRI images. The network only requires inputting the data from a new patient. It achieves high accuracy of classification. It is capable of assisting in diagnosis, leading to early treatment of the different subtypes of sub-cortical ischemia.

Deep learning can also support the evaluation of a semi-continuous measure of the severity of a condition. Along these lines, Finck et al. **(ms. 889808)** improve the estimation of lesion load in multiple sclerosis. They investigate the generalizability of a Generative Adversarial Network (GAN) for synthesizing high-contrast double inversion recovery (DIR) images from lower-contrast T1-weighted and FLAIR MRI images. They also propose the use of uncertainty maps to further enhance their clinical utility. They demonstrate that GAN is capable of synthesizing DIR images with virtual multiple sclerosis lesions that cannot be distinguished from measured real lesions. To this end, they applied an attention module and directed the network's attention toward the lesions. They used data obtained in several imaging centers, thus also demonstrating the generalizability of the model to data obtained in a center whose data were not used for training. The method enhances the automatic counting of multiple sclerosis lesions, which is used as a biomarker of the disease severity and an indicator for the required treatment. Chen M. et al. **(ms. 563097) demonstrate early prediction of measures that are used to diagnose cognitive deficits in very preterm infants**. Up to 40% of very preterm infants ($\leq$32 weeks gestational age) are identified with a cognitive deficit at 2 years of age. Yet, an accurate clinical diagnosis of cognitive deficit cannot be made until 3–5 years of age. Chen et al. obtained diffusion MRI data, computed diffusion-based connectome, and applied transfer learning enhanced deep convolutional neural networks. The performance is superior to that obtained by current methods. Moreover, Chen et al. identified the brain regions most discriminative to the cognitive deficit. The results suggest that deep-learning models can facilitate early prediction of neurodevelopmental outcomes in very preterm infants.

Although deep learning has been increasingly used for neuroimaging image analysis, classification and diagnosis, it has not gained as much ground over standard multivariate pattern analysis (MVPA) techniques in the classification of electroencephalography (EEG). The high dimensionality and large amounts of noise present in EEG data, coupled with the relatively low number of examples (trials) that can be obtained from human subjects are disadvantages for deep learning. **To enable the use of deep learning for MVPA**, Williams et al. **(ms. 491877) present a method of "paired trial classification"** that involves classifying pairs of EEG recordings as coming from the same class or different classes. This makes it possible to significantly increase the number of training examples, through the combinatorics of pairing trials. The final classification is pursued by means of a "dictionary" approach: compare the novel example to a group of known examples from each class. The method can be used as a dataset-specific distance metric that can be extended to novel uses.

Applying deep learning in neuroimaging has become inevitable and this trend is likely to continue in the near future. The studies in this Research Topic show how deep learning can be beneficial to neuroimaging and related modalities across healthy and diseased brains. Combined with recent developments of explainable artificial intelligence and self- or semi-supervised methods, the findings of this Research Topic could be enhanced for even greater impact.

Author contributions

AS wrote the editorial manuscript. All authors contributed to managing the Research Topic and commented on the editorial manuscript. All authors contributed to the article and approved the submitted version.

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.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Feng, X., Provenzano, F. A., Small, S. A. (2022). A deep learning MRI approach outperforms other biomarkers of prodromal Alzheimer's disease. *Alzheimers Res. Therapy* 14, 45. doi: 10.1186/s13195-022-00985-x

Küstner, T., Armanious, K., Yang, J., Yang, B., Schick, F., Gatidis, S. (2019). Retrospective correction of motion-affected MR images using deep learning frameworks. *Magn. Reson. Med.* 82, 1527–1540. doi: 10.1002/mrm.27783

Pizarro, R., Assemlal, H. E., De Nigris, D., Elliot, C., Antel, S., Arnold, D., Shmuel, A. (2019). Using deep learning algorithms to automatically identify the brain MRI contrast: implications for managing large databases. *Neuroinformatics* 17, 115–130. doi: 10.1007/s12021-018-9387-8

Zeng, G., Guo, Y., Zhan, J., Wang, Z., Lai, Z., Du, X., Qu, X., Guo, D. (2021). A review on deep learning MRI reconstruction without fully sampled k-space. *BMC Med. Imaging* 21, 195. doi: 10.1186/s12880-021-00727-9

frontiers in Neuroscience

ORIGINAL RESEARCH
published: 24 January 2020
doi: 10.3389/fnins.2020.00015

Shape Information Improves the Cross-Cohort Performance of Deep Learning-Based Segmentation of the Hippocampus

Irene Brusini[1,2]*, *Olof Lindberg*[2], *J-Sebastian Muehlboeck*[2], *Örjan Smedby*[1], *Eric Westman*[2] *and Chunliang Wang*[1] *for the AddNeuroMed Consortium and the Alzheimer's Disease Neuroimaging Initiative*†

[1] Division of Biomedical Imaging, Department of Biomedical Engineering and Health Systems, KTH Royal Institute of Technology, Stockholm, Sweden, [2] Division of Clinical Geriatrics, Department of Neurobiology, Care Sciences and Society, Karolinska Institute, Solna, Sweden

Performing an accurate segmentation of the hippocampus from brain magnetic resonance images is a crucial task in neuroimaging research, since its structural integrity is strongly related to several neurodegenerative disorders, including Alzheimer's disease (AD). Some automatic segmentation tools are already being used, but, in recent years, new deep learning (DL)-based methods have been proven to be much more accurate in various medical image segmentation tasks. In this work, we propose a DL-based hippocampus segmentation framework that embeds statistical shape of the hippocampus as context information into the deep neural network (DNN). The inclusion of shape information is achieved with three main steps: (1) a U-Net-based segmentation, (2) a shape model estimation, and (3) a second U-Net-based segmentation which uses both the original input data and the fitted shape model. The trained DL architectures were tested on image data of three diagnostic groups [AD patients, subjects with mild cognitive impairment (MCI) and controls] from two cohorts (ADNI and AddNeuroMed). Both intra-cohort validation and cross-cohort validation were performed and compared with the conventional U-net architecture and some variations with other types of context information (i.e., autocontext and tissue-class context). Our results suggest that adding shape information can improve the segmentation accuracy in cross-cohort validation, i.e., when DNNs are trained on one cohort and applied to another. However, no significant benefit is observed in intra-cohort validation, i.e., training and testing DNNs on images from the same cohort. Moreover, compared to other types of context information, the use of shape context was shown to be the most successful in increasing the accuracy, while keeping the computational time in the order of a few minutes.

Keywords: hippocampus, brain MRI, Alzheimer's disease, image segmentation, deep learning, statistical shape model

OPEN ACCESS

Edited by:
John Ashburner,
University College London,
United Kingdom

Reviewed by:
Yi Su,
Banner Alzheimer's Institute,
United States
Jingyun Chen,
New York University, United States

****Correspondence:***
Irene Brusini
brusini@kth.se

†*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*

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *02 May 2019*
Accepted: *08 January 2020*
Published: *24 January 2020*

Citation:
Brusini I, Lindberg O, Muehlboeck J-S, Smedby Ö, Westman E and Wang C (2020) Shape Information Improves the Cross-Cohort Performance of Deep Learning-Based Segmentation of the Hippocampus. Front. Neurosci. 14:15. doi: 10.3389/fnins.2020.00015

INTRODUCTION

Alzheimer's disease (AD) is a chronic progressive neurodegenerative disorder that constitutes approximately 60–70% of all dementia cases (Burns and Iliffe, 2009). The disease is characterized, since its first stages, by the loss of synapses and the depositions of certain lesions in several regions of the brain, which mainly include extracellular Aβ amyloid plaques and intracellular tau neurofibrillary tangles (Vinters, 2015). Moreover, on a macroscopic level, one of the most characteristic signs of the disease is brain atrophy, which is present in the majority of AD patients and can be estimated from magnetic resonance imaging (MRI) (Pini et al., 2016). Therefore, it is important to study imaging biomarkers that could allow early identification of subjects at risk of developing the disorder, as well as quantitatively reflect the disease's level of progression. For example, such biomarkers should be able to distinguish AD both from the healthy state and from mild cognitive impairment (MCI). Indeed, MCI subjects constitute a relevant study group for the early identification of the disease, since several MCI cases, especially when presenting memory dysfunction, have a high probability of later evolving toward AD (Vinters, 2015).

According to the Braak criteria for AD staging (Braak and Braak, 1991), the progression of the disease starts from the transentorhinal cortex (stages I and II), involving then the hippocampus (stages III and IV), and finally spreading to the neocortex (stage V). These steps of progressions were defined based on the changes of accumulation of the neurofibrillary tangles. However, similar patterns can also be seen in the progression of brain atrophy according to multiple MRI studies, which have shown that the atrophy of the hippocampus measured from MRI images can be used, together with the atrophy of the entorhinal cortex, as an early sign of AD (Scheltens et al., 2002). By accurately measuring the volume of these two brain regions, it is possible to separate healthy subjects from AD patients with high precision (Liu et al., 2010). Moreover, shape analysis of the hippocampus has also been shown to be a valid tool for diagnosing AD and differentiating it from other forms of dementia (Lindberg et al., 2012). Evident patterns of hippocampal atrophy have also been reported in several neuroimaging studies on subjects with MCI (Tabatabaei-Jafari et al., 2015).

To properly assess the geometrical features (e.g., volume and shape) of the hippocampus, it is important to have accurate segmentation tools. Ideally, this should be done by completely automated software, since manual segmentation performed by an expert is both extremely time-consuming and relatively subjective. Various software that performs automatic hippocampal segmentation—as well as other brain image processing operations—already exists and is being widely used, for example, in the case of FreeSurfer (Fischl, 2012) or FSL (Jenkinson et al., 2012). However, the computational time of these well-known softwares for performing segmentation is often not acceptable for use in the clinical routine. Moreover, reaching a good segmentation accuracy is a challenging task due to several factors, including, for example, variations in MRI scanners and acquisition modalities, image artifacts, or variations in the brain due to the presence of pathology (Akkus et al., 2017), e.g., hippocampal atrophy.

Several previous studies have explored alternative approaches for automatic brain parcellation. Some of the most popular and successful ways to segment brain MRI images into structures of interest are atlas- and multi-atlas-based segmentation, which consist of integrating information present in brain MRI atlases registered to the target image by using different possible label fusions methods (Cabezas et al., 2011; Asman and Landman, 2013; Wang and Yushkevich, 2013; Pipitone et al., 2014). On the other hand, relevant improvements in the field of medical image segmentation have also been obtained by applying other techniques, such as statistical shape models (Leventon et al., 2000) or the further integration of tissue classifications into multi-atlas-based segmentation (Heckemann et al., 2010). In recent years, very good results have been achieved also by using deep learning (DL)-based methods, which are being more and more widely used because of their superior performance in very diverse medical image segmentation tasks (Ronneberger et al., 2015; Shelhamer et al., 2017). Therefore, such methods—and, in particular, those based on the use of convolutional neural networks—have recently been employed also in several studies on hippocampal segmentation (segmented either alone or together with other brain structures) achieving promising results (Kim et al., 2013; Milletari et al., 2017; Chen et al., 2018; Thyreau et al., 2018).

To further improve the segmentation accuracy, it is general practice to incorporate some context information into the segmentation frameworks. The use of context information, which enables the inclusion of likelihood and priors into the segmentation pipeline, has played an important role in computer vision (Oliva and Torralba, 2007; Tu and Bai, 2010). One example of context information, which has been widely applied in medical image segmentation tasks, is the so-called autocontext. This approach consists of first training one classifier and subsequently using its output as input to a second classifier (Tu and Bai, 2010; Chen et al., 2016; Mirikharaji et al., 2018). Several recent studies have suggested that applying the same strategy to deep neural networks (DNNs) could also improve the segmentation accuracy of brain structures (Chen et al., 2018). Another type of context information can be the tissue-class (Heckemann et al., 2010). More recently, shape context was proposed to help artificial neural networks to segment brain structures (Mahbod et al., 2018). This approach was later extended to DNNs in recent studies that demonstrated how the inclusion of shape priors into the segmentation pipeline can increase the robustness of the network's segmentation accuracy. Such priors were successfully employed, for example, by adding a convolutional autoencoder to a traditional U-Net as shape regularization network (Ravishankar et al., 2017), by feeding a statistical shape model as an additional input to a fully convolutional network (FCN) (Wang and Smedby, 2017), by implementing a Bayesian model that incorporates a shape prior into a DL-based segmentation result (Ma et al., 2018), as well as by jointly training an FCN with a level set (Tang et al., 2017).

In this paper, we investigate whether the integration of shape information can improve the accuracy also in the context of hippocampal segmentation. We analyze the robustness of the method by both testing it on three different diagnostic groups of interest [healthy controls (HCs), MCI subjects, and AD patients] and validating it on a different cohort than the one used for training. Moreover, we compare the effect of adding shape context with two other types of context information: auto-context and tissue-class context. The inclusion of shape information is obtained by building FCNs that receive as input both a T1-weighted MRI image and a statistical shape model of the hippocampus, as already proposed in a previous study on a different segmentation task (Wang and Smedby, 2017). This is done by limiting preprocessing as much as possible, in order to obtain a very fast segmentation (in the order of a few minutes) that could potentially be integrated in the clinical routine.

TABLE 1 | Description of the training and test datasets.

Cohort	Magnetic field strength	HC	AD	MCI	Number of subjects
Training and validation of the network					
ADNI	1.5 T	9	9	9	27
ADNI	3 T	9	9	9	27
Total number of subjects used for training and validation:					54
Only testing					
AddNeuroMed	1.5 T	15	9	13	37
ADNI	1.5 T	213	312	179	704
ADNI	3 T	1799	875	2570	5244
Total number of subjects used only for testing:					5985

The table shows the frequency for the magnetic field strength of the scanners and the subjects' diagnosis in the datasets used for training and testing the proposed hippocampal segmentation pipelines.

MATERIALS AND METHODS

Dataset

For training the networks and validating their performance, 54 T1-weighted structural brain MRI images from the cohort of the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008) were used. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. All ADNI data are obtainable from the ADNI database[1] and updated information is available at www.adni-info.org.

Each of the selected 54 images (of size 197 × 233 × 189, with a voxel size of 1 × 1 × 1 mm^3) had already been manually labeled by experts according to the European AD Consortium and ADNI Harmonized Hippocampal Protocol (HarP) (Boccardi et al., 2015b). The used dataset includes images from scanners of different magnetic field strength (both 1.5 and 3 T) and from three different diagnostic groups: HCs, AD, and MCI. As shown in **Table 1**, the data were selected in such a way that every possible pair of magnetic field strength and diagnosis is represented by the same number of subjects (i.e., nine subjects). No further processing of the images was performed before using them for training the proposed DL pipelines.

Another dataset was then used for further testing the performance of the networks trained only on the above-described ADNI data. It consists of 37 subjects from the AddNeuroMed cohort (Lovestone et al., 2009; Simmons et al., 2009), including all the three analyzed diagnostic groups and acquired using scanners having a magnetic field strength of 1.5 T (see **Table 1**). All 37 MRI images are high-resolution T1-weighted volumes of size of 193 × 229 × 193, with voxel size 1 × 1 × 1 mm. This dataset was chosen because its acquisition protocols were designed in a way compatible with the one used for the ADNI cohort (Simmons et al., 2011; Westman et al., 2011a), so that it is possible to use those data with a DL network previously trained using ADNI data. However, differences in terms of MRI scanner types, image quality, and image size are inevitably present between the two datasets, so it is useful to test the networks on the AddNeuroMed data to check also their performance on images from a new unknown cohort. Moreover, for those 37 MRI images, manual hippocampal segmentations were performed by an expert by following the HarP protocol, so a ground-truth segmentation mask was available.

Finally, the trained networks were tested also on a separate large dataset from the ADNI cohort including 5948 T1-weighted brain images (see **Table 1**). For these data, ground-truth manual segmentation masks of the hippocampus were not available. However, segmentations from FreeSurfer 6.0—processed through TheHiveDB neuroimaging database (Muehlboeck et al., 2013)—could be employed to check their consistency with the result obtained from the DL pipeline.

Segmentation Pipeline

The segmentation methods tested on the data described in the previous section consist of a maximum of three main steps (see **Figure 1**). For each method, a first 3D segmentation is performed using three orthogonal 2D U-Nets which take the original MRI image as input. This first segmentation approach is going to be referred to as *MRI U-Net*.

Moreover, a second segmentation method is presented, which adds a further step to the MRI U-Net. It consists of cropping the original MRI images around both the left and the right hippocampus (preliminarily segmented by using the MRI U-Net) and using the cropped images as input to other three orthogonal U-Nets. This approach is going to be referred to as *Cropped MRI U-Net*.

Finally, we propose a third approach that, after cropping the input MRI images, adds a further step consisting of fitting a statistical shape model to the segmentation obtained from MRI U-Net. Three other orthogonal U-Nets are employed, now taking two images as input: (1) the cropped MRI data and (2) their fitted shape model. This final methodology is going to be referred to as *Shape MRI U-Net*.

MRI U-Net

To perform the first DL-based segmentation, an FCN architecture was implemented: the so-called U-Net, proposed

[1] adni.loni.usc.edu

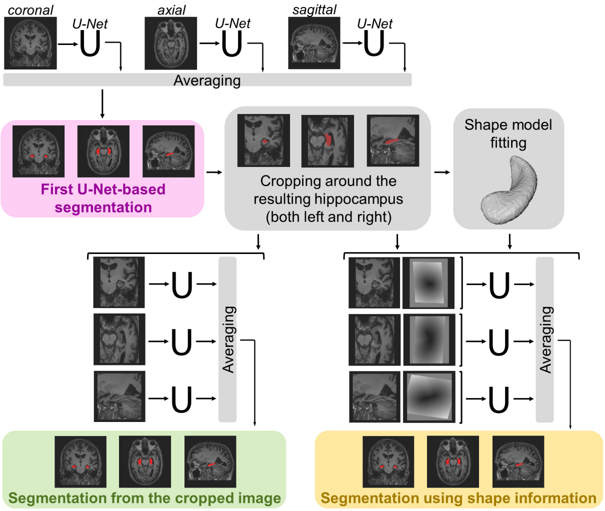

coronal
U-Net
axial
U-Net
sagittal
U-Net
Averaging
First U-Net-based segmentation
Cropping around the resulting hippocampus (both left and right)
Shape model fitting
Averaging
Averaging
Segmentation from the cropped image
Segmentation using shape information

FIGURE 1 | Schematic representation of the implemented segmentation pipelines. Hippocampal segmentation masks are created by applying all the three proposed hippocampal segmentation methods: MRI U-Net (in pink); Cropped MRI U-Net (in green); Shape MRI U-Net (in yellow). All networks receive T1-weighted MRI volumes as inputs, which, in the case of Cropped MRI U-Net and Shape MRI U-Net, are cropped around the hippocampus. Shape MRI U-Net also receives a second input channel that encodes *a priori* hippocampal shape information, which is represented as a distance map from the hippocampal surface (i.e., negative values inside the surface and positive values outside) obtained after a model fitting step.

by Ronneberger et al. (2015), which has been shown to be particularly suitable for medical image segmentation tasks. One of its main strengths is that it can be applied to images of any size, providing as output a probabilistic label map whose dimension is proportional to that of the input image. This is achieved by replacing the fully connected layers of a classical convolutional neural network with more convolutional layers.

Since the segmentation needs to be performed on 3D brain volumes, we implemented three separate U-Nets, which make up the proposed *MRI U-Net* architecture. Each of these U-Net was trained independently to segment 2D slices acquired in one of the three orthogonal views (i.e., axial, coronal, or sagittal). The original T1-weighted image along with the manual segmentation is the only input given to the network for training. The final probability map of the hippocampal segmentation—including both the left and the right hippocampus—is generated by averaging the outputs of the three U-Nets. The final binary segmentation mask (pink box in **Figure 1**) is obtained by taking all voxels having a probability of belonging to either the left or the right hippocampus that is greater or equal to 0.5, i.e., which at least two of the U-Nets agree on classifying as hippocampus.

Cropped MRI U-Net

Once a binary segmentation mask has been obtained from MRI U-Net, it is possible to automatically discriminate the left from the right hippocampus by identifying the two major clusters of voxels in the segmentation mask using the tool "Cluster" from FMRIB Software Library (FSL) (Jenkinson et al., 2012). According to the orientation of the employed images, the right hippocampus is identified as the cluster whose center of gravity has the lowest x coordinate, while the left hippocampus is the remaining cluster.

Once the coordinates of the centers of gravity have been found, it is possible to crop the original MRI image around both the left and right hippocampus. In this way, from each subject, two new 3D volumes are obtained, each having the same predefined size (i.e., 87 × 105 × 111). These cropped volumes are used as input to the three new orthogonal U-Nets, making up the *Cropped MRI U-Net* architecture. Also in this case, the final label map (green box in **Figure 1**) is estimated by averaging the outputs of the three U-Nets and thresholding all voxels having a probability that is greater or equal to 0.5.

Generation of the Shape Model

The volumetric statistical shape models proposed by Leventon et al. (2000) were employed to add the shape context to the DL pipeline. The segmentations used to generate the statistical model were 12 manual labels for the left hippocampus and 12 for the right hippocampus. The total 24 segmentations were obtained from 12 images from the ADNI dataset of 54 subjects. The selection of these images was performed in such a way that all diagnostic groups were equally included (i.e., four HC, four AD, and four MCI), so that the model could represent the variability given by the different diagnoses without overfitting to a specific group. Moreover, all the selected images were acquired from scanners having a magnetic field strength of 3 T, in order to create the model from data of the highest quality available. The choice of the four subjects for each diagnostic group was performed randomly.

Three main modifications were performed on the manual labels. First, each of the 12 images was cropped twice—once around the left and once around the right hippocampus—so that each cropped image only included a region of size 87 × 105 × 111 around one of the regions of interest, similarly to what was done for the input images described in Section "Cropped MRI U-Net." A main advantage of using cropped images also for creating the shape model is the reduction of computational time, since we are not interested in analyzing the rest of the 3D volume that does not include the hippocampus. Second, all labels from the right hippocampi were mirrored in order to match the orientation of the left hippocampus and to create a unique model for both sides together. Finally, each segmentation was up-sampled from 1 to 0.5 mm voxels to improve the resolution of the model. This was done to include more structural details and at the same time avoid large images by limiting the volumetric shape representation to the cropped region.

To generate the model, the mean signed distance function of each of the 24 manually segmented regions was computed, together with five main variations extracted via principal component analysis. Once the model is created, it is possible to fit it to each label map derived from the previous step (presented in section "MRI U-Net") by solving a level set function, as described by Leventon et al. (2000). This fitting step generates a customized hippocampal shape that deviates from the mean shape by adding those variations. The shape fitting step could potentially correct for possible segmentation errors and irregularities present in the MRI U-Net output.

Shape MRI U-Net

The segmentation masks derived from MRI U-Net are cropped around the centers of gravity of both the left and the right hippocampus. Each of these two segmented sides can be associated to its own shape context as described in the previous section. Such shape context consists of a distance map from the hippocampal surface obtained after the model fitting step described in the previous section. In the distance map, the hippocampal surface corresponds to the zero level set, while all voxels inside the surface have negative intensity values and all voxels outside have positive values.

After this, both the cropped original MRI volumes (described in section "Cropped MRI U-Net") and the distance maps from the fitted hippocampal surface are used as inputs to three new U-Nets, which constitute the *Shape MRI U-Net* pipeline. Also in this case, the three networks are trained independently from scratch in the three different orthogonal views. The final segmentation mask (yellow box in **Figure 1**) is estimated by averaging the outputs of the three U-Nets and thresholding the voxels having a probability that is greater or equal to 0.5.

Implementation Details

To build the DL architecture, we used the Keras framework[2]. The implemented U-Nets are identical to those proposed in the original paper by Ronneberger et al. (2015). However, to adapt the images to all the down- and up-sampling steps of the original implementation, all the original brain MRI data (having size 197 × 233 × 189) are resized to 208 × 224 × 192 for the first segmentation step, i.e., MRI U-Net. Instead, for the second (Cropped MRI U-Net) and third (Shape MRI U-Net) approaches, the input data (87 × 105 × 111) are resized to 96 × 112 × 112.

Two different data normalization methods are applied to the input T1-weighted volumes and the shape context images. For the latter ones, the voxel intensity is divided by the standard deviation (computed from all subjects) while keeping the reference point of 0 that corresponds to the hippocampal surface. As for the MRI scans, their intensities are first normalized individually by mapping the lower 5% cutting point of each subject's histogram to 0 and the upper 5% to 1. Afterward, the images are also normalized all together by subtracting the group mean from all subjects and dividing the intensities by the group standard deviation, so that the normalized images have zero mean and a standard deviation of 1.

During the training phase, data augmentation is also employed by generating mini-batches of data in real-time. The data generator randomly applies rotations (within a range of $\pm 10°$), width shifts (range of $\pm 0.1 \cdot$ total image width), height shifts (range of $\pm 0.1 \cdot$ total image height), and zooming (original 100% zoom $\pm$ 20%).

We used the negative Dice score as the loss function to be minimized during the training phase. The number of epochs was always set to 60 for all the three U-Nets of the first pipeline that uses only the T1-weighted volumes as inputs. Instead, for Cropped MRI U-Net and Shape MRI U-Net, the number of epochs was reduced to 40 for all the U-Nets.

Alternative Segmentation Methods

The same images used for training and testing our architecture had also been segmented using the last version (6.0) of FreeSurfer[3], a software tool for image analysis that is freely available online. It is one of the most commonly used tools for automatic segmentation of the subcortical white matter and deep gray matter volumetric structures, including the hippocampus (Fischl et al., 2002, 2004). For this reason, it was chosen as a reference method for comparison of the performance of the pipeline proposed in our work.

To better investigate the contribution given by adding the shape-model-fitting step, we compared the performance of our proposed pipelines with two alternative types of context information too. They were integrated in the following two networks:

Tissue MRI U-Net

The hippocampus is a gray matter structure that borders with other tissue types in specific locations, so *a priori* tissue type classification could help the network to identify the boundaries of the hippocampus. Therefore, three main tissue type segmentations (i.e., gray matter, white matter, and cerebrospinal fluid) are used as context input to the network to be integrated to the cropped MRI image. Such segmentations could be obtained automatically from the original MRI image by using the FMRIB's Automated Segmentation Tool (FAST) from FSL (Zhang et al., 2001; Jenkinson et al., 2012). The network thus has four input channels in total: the cropped MRI image (obtained in the same way as in Cropped MRI U-Net and Shape MRI U-Net), as well as the three tissue segmentations (cropped in the same location as the T1-weighted images).

Autocontext MRI U-Net

The autocontext strategy is used: the cropped segmentation derived from *MRI U-Net* is given as second input channel to the network. This technique is one of the most well-known types of context information (Tu and Bai, 2010; Chen et al., 2016; Mirikharaji et al., 2018) and we aimed at investigating its effect also on our application of interest.

Method Evaluation

Single-Cohort Evaluation

The proposed methods were first evaluated using ninefold cross-validation on the first dataset of 54 subjects from the ADNI cohort, for which manual hippocampal segmentations were available and used during training. For each fold, 48 of the cases were used for training and the remaining six for testing, and the test set always included all the six possible combinations of magnetic field strength and diagnosis presented in **Table 1**. When the shape context was included in the pipeline, the shape model

[2] https://keras.io

[3] http://surfer.nmr.mgh.harvard.edu/

described in Section "Generation of the Shape Model" was always the same and not re-created for each fold.

The evaluation metrics used to analyze the accuracy were the Dice score, precision, recall, and Hausdorff distance. The Dice score (Dice, 1945) is an index that measures the degree of overlap between two segmentation masks with values between 0 (no overlap) and 1 (matching segmentations). When comparing a hippocampal segmentation result with its ground truth, we can define as "true positives" (TP) the number of voxels correctly classified as belonging to the hippocampus and "false positives" (FP) those wrongly classified as belonging to the hippocampus. On the other hand, voxels correctly classified as background are "true negatives" (TN) and those wrongly classified as background are "false negatives" (FN). Given these definitions, we could estimate the Dice score as $\frac{2\,TP}{2TP+FP+FN}$, the precision as $\frac{TP}{TP+FP}$, and the recall as $\frac{TP}{TP+FN}$. Finally, as regards the Hausdorff distance, since it is a metric that tends to be highly sensitive to the presence of outliers (Huttenlocher et al., 1993; Taha and Hanbury, 2015), we applied the quantile method proposed by Huttenlocher et al. (1993). This method consists of, first of all, computing the closest distance between every point of a segmentation mask and the ground-truth. After this, these computed distances are sorted (from the lowest to the highest) and, instead of simply identifying their maximum value (i.e., the classical Hausdorff distance), their qth quantile is reported. In particular, in this paper, the 95th percentile of the distances was computed for each subject.

Cross-Cohort Evaluation

Once the segmentation pipelines had been trained and validated as described above, they were tested on a new unseen dataset of 37 subjects from the AddNeuroMed cohort. Since ground-truth segmentation masks were available for this dataset, the same evaluation metrics described in the previous section were used. Moreover, the performance differences between the methods were analyzed by carrying out pairwise comparisons between the evaluation metrics obtained from all the tested DL-based pipelines. This was done by defining a mixed-effect analysis of variance model with the segmentation methods and the sides (i.e., left or right) as fixed effects, the subjects as random effect, and the resulting evaluation metrics as dependent variables. The statistical calculation was performed in Stata 13.1 (StataCorp, College Station, TX, United States).

Finally, the same networks were also trained on the AddNeuroMed dataset and tested on the ADNI dataset, which had previously been used as training set. Dice score, precision, recall, and Hausdorff distance were computed. This was done in order to further investigate the effect of training the pipelines on data from a certain cohort and testing them on a new unseen cohort.

Evaluation on a Larger ADNI Dataset

As a final test of the trained networks, the proposed DL-based segmentation methods, which were trained on the above-described ADNI dataset of 54 subjects, were also applied on a separate larger dataset of 5948 T1-weighted brain images still from the ADNI cohort. Given the large amount of data and the consequently long time needed to perform all the segmentations, we only tested two pipelines in this phase: one not including shape information—MRI U-Net—and one including such information—Shape MRI U-Net.

For this dataset, ground-truth manual segmentation masks of the hippocampus were not available, but segmentations from FreeSurfer 6.0 could be easily obtained. Therefore, they were employed to check their consistency with the result obtained from the DL pipelines. This choice was motivated by the fact that FreeSurfer is still among the most commonly used software for brain image analysis. Thus, a similarity between our results and those from FreeSurfer could allow us to the test the potential of our methods to possibly replace one tool that is already well-known and established. Such consistency was analyzed by computing the correlation between the FreeSurfer volumes and those obtained through each DL-based pipeline. Moreover, two additional shape similarity metrics (i.e., Dice score and Hausdorff distance) were also computed to evaluate the similarity between the results from the proposed pipelines and those from FreeSurfer.

In addition, given the large amount of data in this dataset, we investigated whether our segmentation results could reflect the volumetric changes in the hippocampus between the three diagnostic group of interest, i.e., AD, MCI, and HC. This was done by selecting all subjects whose diagnosis did not change within 2 years after the first MRI scan. Subsequently, we computed, for each subject, the hippocampal volume at baseline and divided it by the total intracranial volume (ICV), which is one of the outputs measurements (eTIV, estimated total ICV) given by FreeSurfer 6.0. This normalization is often used in literature in order to have a more reliable estimation of atrophy caused by neurodegeneration (Voevodskaya et al., 2014). This measurement was then multiplied by the average ICV for all the subjects of interest. A one-way ANOVA test was then employed to identify whether a statistically significant difference ($p < 0.05$) could be found between the groups. Moreover, the normalized volumes were also used to fit three binary logistic regression models (i.e., for AD vs. HC, AD vs. MCI, and MCI vs. HC) to investigate the possibility of predicting the diagnosis of one subjects from the above-described volumetric measurements. In particular, each model was generated to provide as output the probability of a subject to belong to a certain diagnostic group as a function of the hippocampal volume multiplied by the ratio between the ICV and the specific subject's ICV. The prediction power of each binary model was analyzed by computing three evaluations metrics: area under the curve (AUC), sensitivity, and specificity.

Finally, in this dataset, 2704 of the scans were repeated twice on the same subject, with the same scanner and at the same time point (within the same week). This allowed us to perform a test–retest analysis to make sure that the implemented methods are reproducible and consistent between the two subsequent scans. Therefore, for each of the tested methods, we computed the concordance correlation coefficient (CCC) (Lin, 1989) between the hippocampal volumes from the two subsequent scans. This coefficient describes the agreement between two different measurements of the same variable. CCC varies between –1 and 1, and CCC = 1 indicates perfect reproducibility.

RESULTS

Single-Cohort Evaluation

The performance of the three types of context-aware segmentation methods was evaluated on the first ADNI dataset of 54 subjects through ninefold cross-validation and compared between the preliminary segmentation step (MRI U-Net) and the additional steps using only cropped data (Cropped MRI U-Net) or cropped data together with shape context (Shape MRI U-Net), as shown in **Table 2**. No relevant differences were found between the three methods, which showed quite consistent results on both the left and the right hippocampus. Among all the tested DL-based methods, Tissue MRI U-Net showed the worst performance, having a slightly lower accuracy and higher Hausdorff distance in average compared to the other methods.

The average Dice score was also estimated within each of the three diagnostic groups (HC, AD, and MCI). This was done to check whether the system has a consistent performance across all possible forms of hippocampal integrity. As shown in **Figure 2**, all diagnostic groups showed a similar segmentation accuracy in both the left and right hippocampus by using the three proposed methods. However, the AD patients always presented a slightly lower Dice score (1 or 2% lower in average) with respect to the other two subject groups.

As presented in **Table 2**, our methods yielded better values than FreeSurfer in all the considered evaluation metrics. This applies also for the comparison between diagnostic groups (see **Figure 2**), in which, contrary to the proposed DL methods, FreeSurfer showed a higher performance loss when dealing with MCI and—even more—AD subjects, compared to the HCs.

In order to better understand the influence of each of the three independent U-Nets (one for each view) toward the final segmentation, we also computed the evaluation metrics separately for each U-Net (see **Supplementary Table S1**). These results showed how, for MRI U-Net, the highest accuracy is obtained on the axial view. By contrast, for Cropped MRI U-Net and Shape MRI U-Net, the highest accuracy can be observed on the coronal view. However, while no big differences can be found across all views in MRI U-Net and Shape MRI U-Net, Cropped MRI U-Net showed an evident decrease in performance on the sagittal view in terms of Dice score, precision, and Hausdorff distance.

Moreover, the present methods were proven to be more efficient also in terms of computational time, at least when only a hippocampal segmentation is desired. On a personal computer with an Nvidia GTX 1080 graphic card and 32 GB of RAM, each segmentation took between 25 and 30 s with the simple MRI U-Net methodology. When performing one segmentation with Cropped MRI U-Net, approximately 1 min was taken. Using Shape MRI U-Net, about two and a half minutes was needed for one subject.

Cross-Cohort Evaluation

Testing on the AddNeuroMed Dataset

When the proposed segmentation pipelines were tested on the new unseen dataset from the AddNeuroMed cohort, larger differences between the tested methods could be observed (see **Table 3**).

The accuracy achieved by segmenting the MRI images using the MRI U-Net architecture is now very close to that obtained by using FreeSurfer 6.0. In particular, the two methods have almost identical Dice scores, while the precision and recall are, respectively, decreased and increased by using MRI U-Net. Moreover, FreeSurfer has a slightly higher Hausdorff distance in average.

When performing the segmentation using the other two proposed pipelines, an improvement in the performance can be observed. Dice score, precision, and recall positively increased by using the Cropped MRI U-Net architecture and, even more, the Shape MRI U-Net. The benefit of adding shape context was particularly noticed in the right hippocampus, where the average Dice score increased by 4.04% with respect to MRI U-Net

TABLE 2 | Single-cohort evaluation.

Region of interest	Segmentation method	Dice score	Precision	Recall	Hausdorff distance (in voxels)
Left hippocampus	MRI U-Net	90.17 ± 1.44%	89.46 ± 2.20%	90.96 ± 2.29%	2.33 ± 0.55
	Cropped MRI U-Net	90.28 ± 1.30%	89.36 ± 2.20%	91.28 ± 1.85%	2.22 ± 0.53
	Shape MRI U-Net	90.01 ± 1.41%	88.55 ± 2.69%	91.60 ± 1.87%	2.35 ± 0.67
	Tissue MRI U-Net	88.79 ± 1.61%	86.79 ± 2.72%	91.01 ± 2.94%	2.39 ± 0.60
	Autocontext MRI U-Net	89.45 ± 1.46%	86.69 ± 2.77%	92.53 ± 2.75%	2.30 ± 0.57
	FreeSurfer 6.0	79.52 ± 3.14%	82.94 ± 5.01%	76.60 ± 3.94%	4.34 ± 1.08
Right hippocampus	MRI U-Net	90.12 ± 1.41%	89.59 ± 2.48%	90.77 ± 2.72%	2.39 ± 0.53
	Cropped MRI U-Net	90.26 ± 1.41%	89.29 ± 2.62%	91.35 ± 2.39%	2.47 ± 0.59
	Shape MRI U-Net	90.08 ± 1.67%	88.50 ± 3.39%	91.86 ± 2.34%	2.54 ± 0.79
	Tissue MRI U-Net	88.74 ± 1.50%	86.4 ± 2.84%	91.00 ± 3.16%	2.63 ± 0.70
	Autocontext MRI U-Net	89.63 ± 1.32%	87.25 ± 2.81%	92.30 ± 2.89%	2.39 ± 0.53
	FreeSurfer 6.0	80.21 ± 3.86%	83.63 ± 4.35%	77.31 ± 5.36%	4.50 ± 1.23

The performance of the proposed methods (in terms of Dice score, precision, recall, and Hausdorff distance) was computed through ninefold cross validation and compared with that of FreeSurfer 6.0. All evaluation metrics are expressed as mean ± standard deviation.

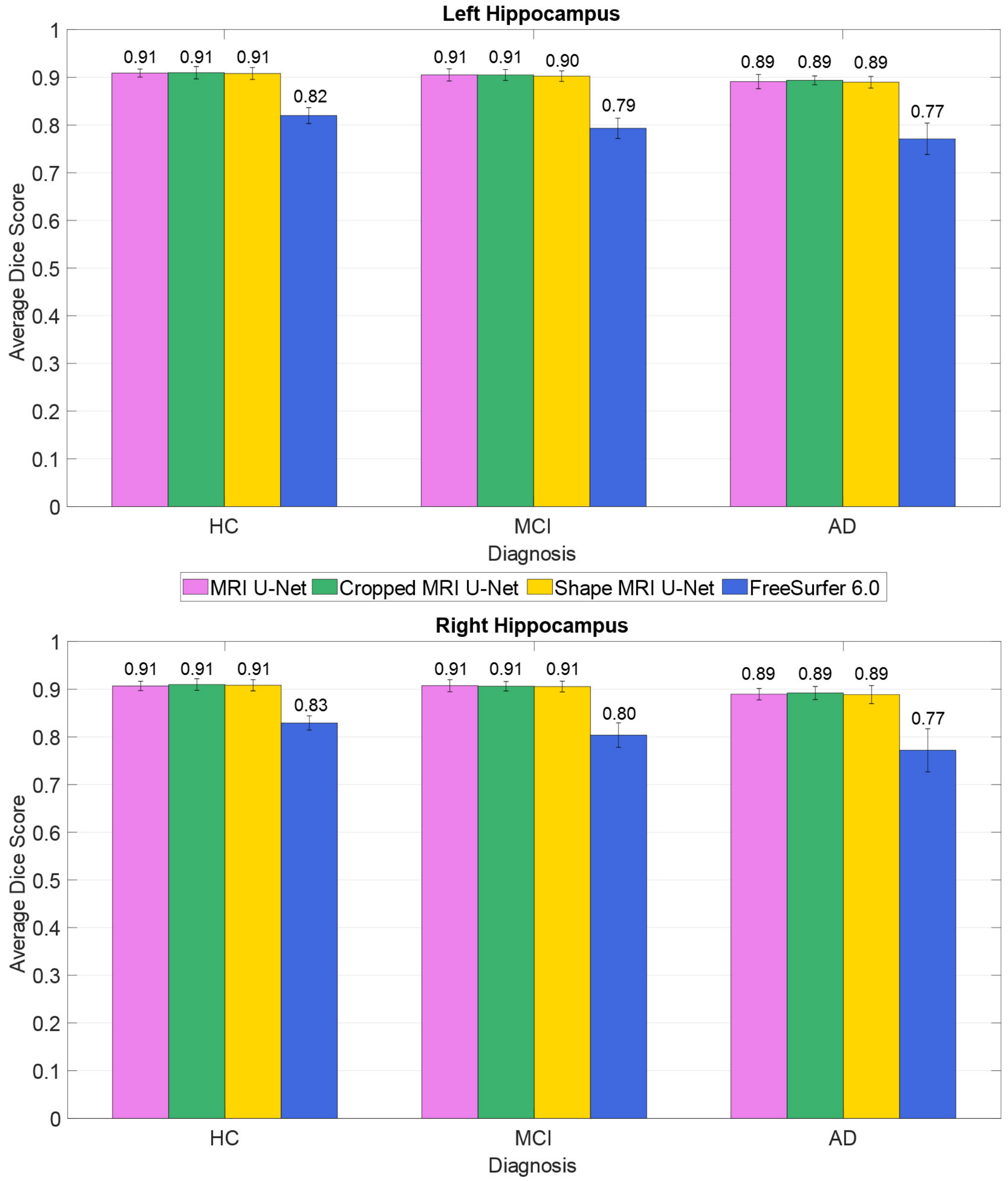


FIGURE 2 | Difference in segmentation accuracy (from cross validation) between the three analyzed diagnostic groups (HC, MCI, and AD). The accuracy is expressed as the Dice score averaged across all subjects and is represented with histograms for each of the tested methods (see color legend). The error bars show the standard deviation of the Dice score.

(compared to + 2.70% obtained with Cropped MRI U-Net), the average precision by 3.40% (compared to + 1.43%), and the average recall by 5.03% (compared to + 4.47%). For the left hippocampus, the difference between Cropped MRI U-Net and Shape MRI U-Net was less evident, but in both cases, all evaluation metrics increased by approximately 4% with respect to MRI U-Net.

Also for this analysis, we calculated the evaluation metrics separately for each independent 2D U-Net (see **Supplementary Table S2**). Similarly to what has been obtained for the single-cohort analysis, MRI U-Net showed its best accuracy on the axial input slices, while with Cropped MRI U-Net and Shape MRI U-Net no relevant differences could be noticed between coronal and axial views. Moreover, most of the single 2D U-Nets of Cropped MRI U-Net showed a lower performance compared to Shape MRI U-Net, whose results are also more consistent across views. In particular, the sagittal 2D U-Net of Cropped MRI U-Net was still shown to have a very high Hausdorff distance compared to all other views and approaches, as well as particularly low Dice score and precision.

The two alternative integrations of context information did not achieve a better performance than the proposed methods. In

TABLE 3 | First cross-cohort evaluation.

Region of interest	Segmentation method	Dice score	Precision	Recall	Hausdorff distance (in voxels)
Left hippocampus	MRI U-Net	79.09 ± 2.63%	74.72 ± 4.27%	84.23 ± 3.15%	3.44 ± 0.74
	Cropped MRI U-Net	84.44 ± 2.32%	78.47 ± 4.17%	91.60 ± 2.47%	3.19 ± 0.64
	Shape MRI U-Net	84.92 ± 2.56%	79.46 ± 5.03%	91.57 ± 3.60%	3.16 ± 0.77
	Tissue MRI U-Net	84.32 ± 2.16%	79.04 ± 4.12%	90.59 ± 2.90%	3.33 ± 0.85
	Autocontext MRI U-Net	80.55 ± 2.61%	73.99 ± 4.23%	88.67 ± 3.50%	3.33 ± 0.72
	FreeSurfer 6.0	79.41 ± 3.77%	78.89 ± 5.46%	80.20 ± 4.39%	4.24 ± 1.25
Right hippocampus	MRI U-Net	80.15 ± 2.25%	74.54 ± 3.12%	86.80 ± 3.08%	3.92 ± 1.14
	Cropped MRI U-Net	82.85 ± 2.52%	75.97 ± 3.91%	91.27 ± 2.31%	3.80 ± 1.05
	Shape MRI U-Net	84.19 ± 2.50%	77.94 ± 4.49%	91.83 ± 3.28%	3.62 ± 1.04
	Tissue MRI U-Net	82.88 ± 2.35%	76.86 ± 3.60%	90.08 ± 2.71%	3.68 ± 1.11
	Autocontext MRI U-Net	80.51 ± 2.20%	73.08 ± 3.27%	89.79 ± 3.03%	3.88 ± 1.16
	FreeSurfer 6.0	79.57 ± 3.54%	77.71 ± 5.53%	81.78 ± 3.40%	4.61 ± 1.11

The performance of the proposed methods (in terms of Dice score, precision, recall, and Hausdorff distance) was tested on a new unseen dataset from a different cohort (i.e., AddNeuroMed cohort) than the one used for training. The performance is reported also for the segmentations obtained using FreeSurfer 6.0 on the same data. All evaluation metrics are expressed as mean ± standard deviation.

particular, Autocontext MRI U-Net showed a very similar result to MRI U-Net. Instead, with Tissue MRI U-Net, the performance is comparable to that of Cropped MRI U-Net and Shape MRI U-Net, but never outperforming them in any of the analyzed evaluation metrics.

As shown in **Supplementary Table S3**, a statistically significant difference (i.e., $p < 0.05$ with Bonferroni correction) was found in the majority of the pairwise comparisons between the tested segmentation methods for all the evaluation metrics, except for the Hausdorff distance. The value of this latter metric is indeed quite consistent across all DL-based methods (except for the difference between Shape MRI U-Net and MRI U-Net, which resulted to be significant). Significantly larger Hausdorff distances were, however, always found when using FreeSurfer 6.0 as opposed to the pipelines implemented in the present work. Moreover, the choice of the subject to be segmented—and, subsequently, the image quality, as well as the level of degeneration—was found to highly influence the performance. **Figure 3** shows how, for each subject, the evaluation metrics tended to vary with a consistent pattern according to the method being used and maintained a rather similar between-subject variability within each method.

The average Dice scores for each diagnostic group from the AddNeuroMed dataset were also analyzed, as shown in **Figure 4**. The results reflect what has been observed on the whole dataset (i.e., averaging the results across all 37 subjects): both Cropped MRI U-Net and Shape MRI U-Net showed a superior accuracy compared to MRI U-Net, and in general the DL-based methods performed better than FreeSurfer 6.0.

Testing on the ADNI Dataset

All the implemented networks were re-trained using the data from the AddNeuroMed cohort in order to be tested on the 54 subjects from the ADNI cohort that had previously been used for training. Average Dice score, precision, recall, and Hausdorff distance were computed and presented in **Table 4**. The results are rather consistent with what has been found for the first cross-cohort evaluation presented in Section "Testing on the AddNeuroMed Dataset." The average Dice scores for MRI U-Net and Autocontext MRI U-Net are very similar to those obtained using FreeSurfer 6.0, while they increase when using Cropped MRI U-Net and Shape MRI U-Net. However, in this case, the best results in terms of Hausdorff distance could be found in Autocontext MRI U-Net, followed by the Shape MRI U-Net implementation.

Two major differences could be found compared to the previous cross-cohort evaluation. First, Tissue MRI U-Net showed a much worse performance in terms of Dice score, precision, and recall. Second, all the other deep-learning based methods resulted in having both higher precision and lower recall compared to the previous analysis.

Testing on a Larger ADNI Dataset

The 5948 additional cases from the ADNI cohort were segmented using the networks trained on the above-described balanced ADNI dataset of 54 subjects. The correlation coefficients of the volumetric results were rather high and consistent between each of the two tested pipelines and FreeSurfer, as can be observed in **Figure 5**. For the sake of completeness, we also computed the correlation between the two present U-Net based pipelines as well, which resulted in a correlation coefficient of 0.952 for the left and 0.958 for the right hippocampus. Thus, there is a higher correlation between the tested DL-based pipelines than between either of these pipelines and FreeSurfer.

The scatter plots of **Figure 5** highlight the presence of a few outliers, whose number appears to be higher using MRI U-Net but decreases with Shape MRI U-Net. For each of the proposed pipelines, we computed the hippocampal volume of every subject—obtained after applying one of the given segmentation pipelines—divided by the hippocampal volume obtained, instead, from FreeSurfer on the same subject. These ratios were then used to extract a measure of the amount of outliers. We defined as outliers all those subjects that, for a specific segmentation pipeline, showed a volumetric ratio deviating from the median ratio by at least three times the median absolute deviation. The

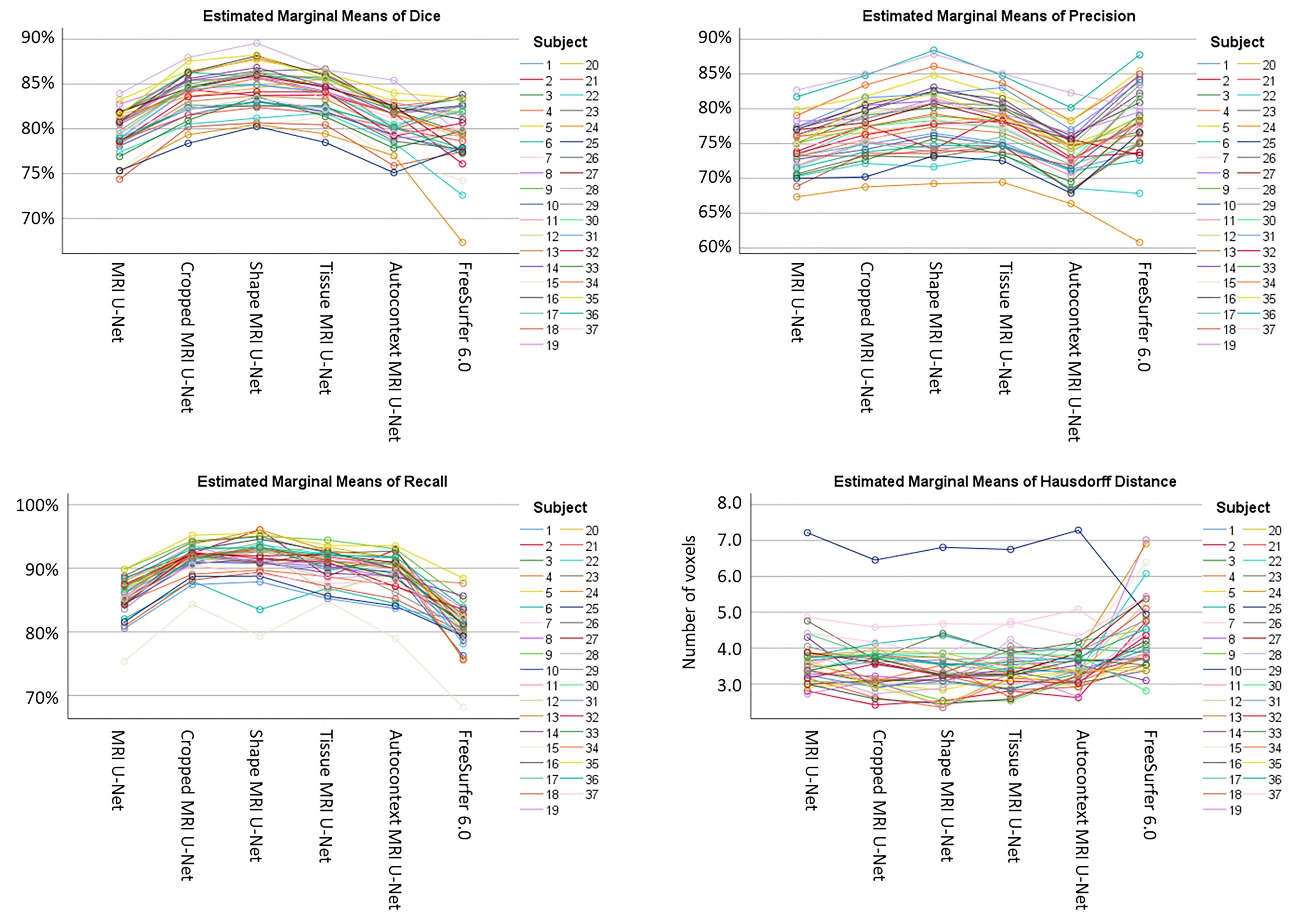


FIGURE 3 | Variability of the performance within subjects on the AddNeuroMed test dataset. The plots are showing how, for each subject (see color legend), the evaluation metrics (on the vertical axes) change according to the segmentation method (on the horizontal axes) being used. Each of the four plots represents one specific evaluation metric: Dice score (top left), precision (top right), recall (bottom left), and Hausdorff distance (bottom right).

results confirmed what could be seen from the plot. Indeed, for the left and right hippocampus, respectively, 104 and 140 outliers were identified using MRI U-Net, while 84 and 96 using Shape MRI U-Net. Of these subjects, 14 appeared as outliers (for both left and right hippocampus) in all three pipelines. All these 14 cases were either MCI subjects or AD patients and examples of the segmentation results in some of those are shown in **Figure 6**. An expert was asked to compare the segmentations obtained from FreeSurfer with those from MRI U-Net (which, as in **Figure 6**, was chosen as reference DL-based segmentation method for this evaluation) in these 14 subjects. In all 14 cases, FreeSurfer showed segmentation errors. With MRI U-Net, instead, three out of these 14 cases showed good segmentation results, five out of 14 showed inaccurate but better results than FreeSurfer, while the remaining six cases were classified as segmentation errors in the same manner as FreeSurfer. Moreover, in **Figure 5**, the plot for the left hippocampal segmentation using MRI U-Net and both plots for the right hippocampus show one specific point that has a very low volume (in some cases very close to zero). This point corresponds to the same subject in all of these three cases. The original MRI scan of this subject was visually inspected, and it was found to be affected by artifacts that made the identification of the hippocampus particularly challenging. The result obtained on the same subjects on the left hippocampus using Shape MRI U-Net was also inaccurate, even if characterized by a larger amount of voxels.

Two additional similarity metrics (i.e., Dice score and Hausdorff distance) have been computed to compare the results from FreeSurfer with those from both MRI U-Net and Shape MRI U-Net (see **Supplementary Table S4**). These results showed a rather high consistency between these methods, with an average Dice score close to 79% for the comparison with MRI U-Net, and around 82% for Shape MRI U-Net. Also the Hausdorff distance was rather low (i.e., around 4 voxels in average) for all methods.

We also investigated whether there is a statistically significant difference in the normalized hippocampal volume between the three diagnostic groups of interest, i.e., AD, MCI, and HC. All the three analyzed segmentation methods (MRI U-Net, Shape MRI U-Net, and FreeSurfer 6.0) resulted in statistically significant differences between all three diagnostic groups. As can be seen in **Table 5**, the lowest normalized hippocampal volumes were always found in the AD patients, and the highest

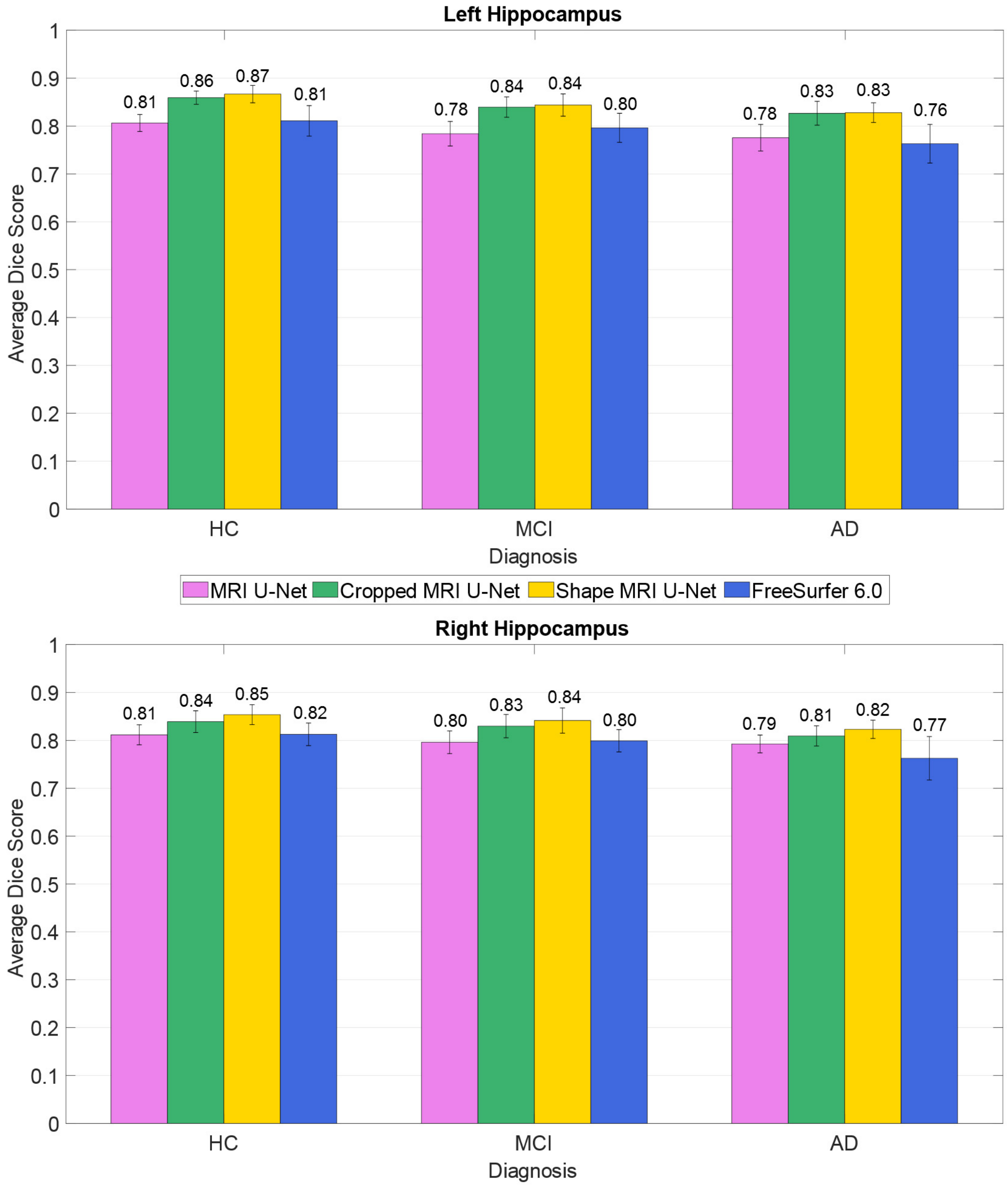


FIGURE 4 | Difference in segmentation accuracy (on the AddNeuroMed test dataset) between the three analyzed diagnostic groups (HC, MCI, and AD). The accuracy is expressed as the Dice score averaged across all subjects and is represented with histograms for each of the tested methods (see color legend). The error bars show the standard deviation of the Dice score.

in the HCs. We then investigated the diagnostic prediction power by computing the AUC, sensitivity, and specificity of three logistic regression models that were fitted to classify AD vs. HC, AD vs. MCI, and MCI vs. HC by using the above-mentioned normalized measurements. The results, which are reported in **Table 6**, show that, for all three segmentations methods, a rather good prediction power is achieved when comparing AD subjects and HC, with a AUC that is above 0.80. Instead, the task of distinguishing AD from MCI and MCI from HC subjects is more challenging, with an AUC of 0.68 for all three methods in the classification of AD vs. MCI and slightly lower AUCs for the classification of MCI vs. HC. Sensitivity and specificity measurements are also shown to be consistent with the AUC across methods and classification tasks. Moreover, the DL-based methods have also shown to have a slightly higher performance compared to FreeSurfer, given their overall higher evaluation metrics.

Finally, in this dataset, we computed the CCCs between the hippocampal volumes from all the available pairs of subsequent test–retest scans from the same subject at the same time point. For the left and right hippocampus, respectively, the CCC resulted in 0.988 and 0.977 with Shape MRI U-Net,

TABLE 4 | Second cross-cohort evaluation.

Region of interest	Segmentation method	Dice score	Precision	Recall	Hausdorff distance (in voxels)
Left hippocampus	MRI U-Net	80.26 ± 3.93%	87.92 ± 3.72%	74.03 ± 5.48%	3.52 ± 0.85
	Cropped MRI U-Net	84.56 ± 2.45%	88.12 ± 2.77%	81.42 ± 4.00%	3.44 ± 0.82
	Shape MRI U-Net	85.06 ± 2.47%	87.85 ± 3.08%	82.57 ± 3.72%	3.34 ± 0.74
	Tissue MRI U-Net	73.39 ± 8.93%	75.66 ± 8.90%	71.40 ± 9.45%	4.30 ± 1.04
	Autocontext MRI U-Net	79.64 ± 7.50%	77.37 ± 7.84%	82.14 ± 7.56%	3.00 ± 0.75
	FreeSurfer 6.0	79.52 ± 3.14%	82.94 ± 5.01%	76.60 ± 3.94%	4.34 ± 1.08
Right hippocampus	MRI U-Net	82.00 ± 3.42%	90.42 ± 2.99%	75.28 ± 5.54%	3.79 ± 0.74
	Cropped MRI U-Net	85.62 ± 1.92%	88.56 ± 2.93%	83.03 ± 3.68%	3.54 ± 0.80
	Shape MRI U-Net	86.06 ± 2.01%	88.20 ± 3.46%	84.21 ± 3.70%	3.47 ± 0.88
	Tissue MRI U-Net	73.59 ± 6.64%	75.42 ± 6.97%	72.03 ± 7.30%	4.19 ± 0.88
	Autocontext MRI U-Net	79.24 ± 6.07%	77.19 ± 6.41%	81.52 ± 6.37%	3.03 ± 0.60
	FreeSurfer 6.0	80.21 ± 3.86%	83.63 ± 4.35%	77.31 ± 5.36%	4.50 ± 1.23

The proposed pipelines were re-trained on the dataset from the AddNeuroMed cohort and tested on the data from the ADNI cohort, which were previously used for training. The performance of the methods is presented in terms of Dice score, precision, recall, and Hausdorff distance. The performance is reported also for the segmentations obtained using FreeSurfer 6.0 on the same data. All evaluation metrics are expressed as mean ± standard deviation.

0.989 and 0.986 with MRI U-Net, and in 0.969 and 0.963 with FreeSurfer 6.0.

DISCUSSION

Comparison Between the Implemented Pipelines

In this work, three different U-Net based segmentation pipelines were proposed: MRI U-Net, Cropped MRI U-Net, and Shape MRI U-Net. All three methods were shown to be accurate and quick tools for the automatic segmentation of the hippocampus from brain MRI data.

Single-Cohort Analysis

The first presented method, MRI U-Net, constitutes the simplest architecture, which takes the original MRI image as input and performs the segmentation using three orthogonal U-Nets. When testing its performance through cross validation on 54 subjects from the ADNI dataset, it was shown to achieve an excellent accuracy (average Dice score of approximately 90%) which was equal to each of the other two proposed and more elaborate methods (Cropped MRI U-Net and Shape MRI U-Net). It also yielded higher accuracy than the software FreeSurfer. To some extent, this was expected since the segmentation protocol of the ground-truth masks coincides with that used to train the network, while it inevitably differs from the atlas on which the FreeSurfer segmentation is based (Fischl et al., 2002). However, given the very high difference in performance between the two methods (i.e., around 10% of improvement in the Dice score), we believe that such comparison is valuable and worth being reported in order to give a measure of how DL-based methods are outperforming older—but still widely used and established—brain image processing software.

These results suggest that, when the training and test set come from the same cohort, the use of the simple T1-weighted scan as input image is more efficient than both using just a portion of the scan (cropped around the hippocampus) and including context information. The step of cropping the image around the hippocampus is probably not needed for the network to increase its performance because data from the same cohort have the same size and very similar scanning quality, and therefore the localization and size of the hippocampal region is quite consistent across images. As regards the lack of improvement by adding shape context layers, it is probably due to the fact that a high accuracy can already be reached by using the preliminary single-channel networks and, as already observed in a previous study (Wang and Smedby, 2017), the inclusion of shape information is most valuable when the structure to be segmented is rather challenging.

The analysis of each independent U-Net (i.e., trained for each view separately) was also useful to better understand the differences between the three approaches, which, globally, seem to be very similar to each other. The coronal view is typically the most used view to perform manual hippocampal segmentation. However, its morphological details are not always sufficient to achieve an accurate results, so the axial and sagittal views have to be checked as well (Boccardi et al., 2015a). Therefore, a superior performance on the coronal view was expected on all the trained U-Nets. However, in MRI U-Net, the best performing network was shown to be the one trained on axial slices, suggesting that this model is able to capture some important image features that differ from those used by the human raters. On the other hand, both Cropped MRI U-Net and Shape MRI U-Net showed a slightly superior performance on the coronal view, which is more consistent with what happens in practice when the segmentation is performed by expert radiologists. Moreover, Cropped MRI U-Net resulted in a relevantly low performance on the sagittal view compared to all other views. In particular, the high average Hausdorff distance suggests the presence of several geometric errors, which are then corrected by integrating the information from the other two views. This could not be observed on Shape MRI U-Net, suggesting that the use of shape information on the sagittal view can help to prevent the occurrence of such geometric errors.

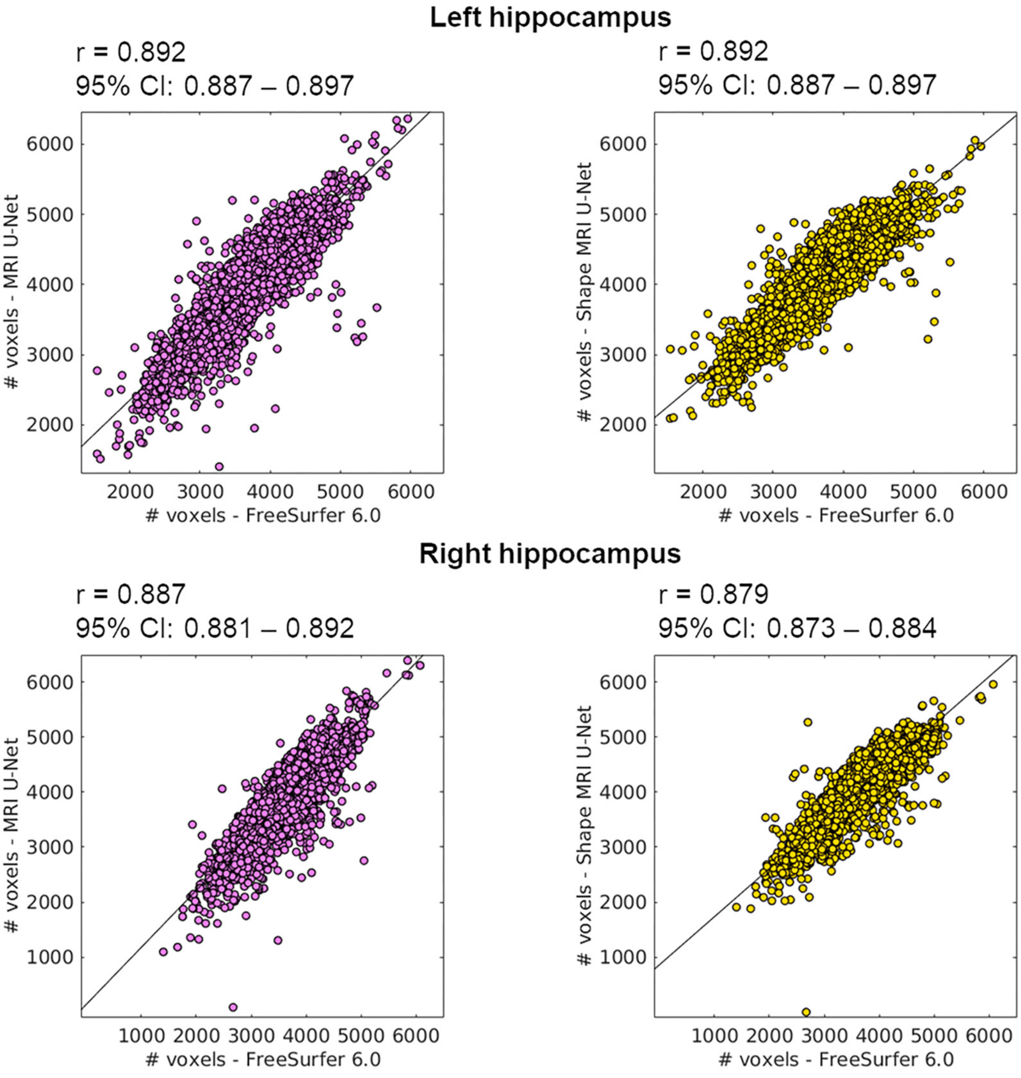


FIGURE 5 | Correlation between the volume of the segmentations obtained using the proposed methods (on the large ADNI test dataset) and those from FreeSurfer 6.0. The Pearson correlation coefficient (*r*) is reported together with its 95% confidence interval (95% CI). Results are reported for both the left (top row) and right (bottom row) hippocampus, and for both MRI U-Net (pink) and Shape MRI U-Net (yellow) in comparison with FreeSurfer 6.0. The volumes are plotted and expressed in terms of number of voxels in the region of interest.

Cross-Cohort Analysis

When the networks trained on the ADNI cohort were tested on a dataset from the AddNeuroMed cohort, the observed differences between the three implemented architectures were subject to a consistent change.

In terms of overall accuracy, MRI U-Net and FreeSurfer, which both process the data by receiving as input only the original T1-weighted image, showed a very similar performance. The main difference between them is a lower precision and higher recall obtained, in average, by using MRI U-Net. The lower precision may be due to an over-estimation of the hippocampal mask in regions where hippocampal atrophy is present. This suggests the difficulties of training a network with enough atrophic patterns to be able to obtain accurate segmentations also on new unseen data. On the other hand, the under-estimations obtained by FreeSurfer may be related to other types of segmentation errors in atrophic hippocampal areas as well, as suggested also by the general decrease in performance in MCI and AD subjects (**Figure 4**). This issue will be subject to future investigations.

Furthermore, a clearly higher accuracy was now observed by employing Cropped MRI U-Net and, even more, Shape MRI U-Net. Therefore, when segmenting new unseen data that differ from those used during training (for example, in terms of image size, scanner types, and image quality), it seems to be motivated to perform a further processing step adding information to the simple MRI scan. A big improvement in the accuracy was seen already by simply cropping the image around the center of gravity of the preliminary hippocampal segmentation, suggesting that already this step largely harmonizes the input images to those used during training. This could be explained by the fact that

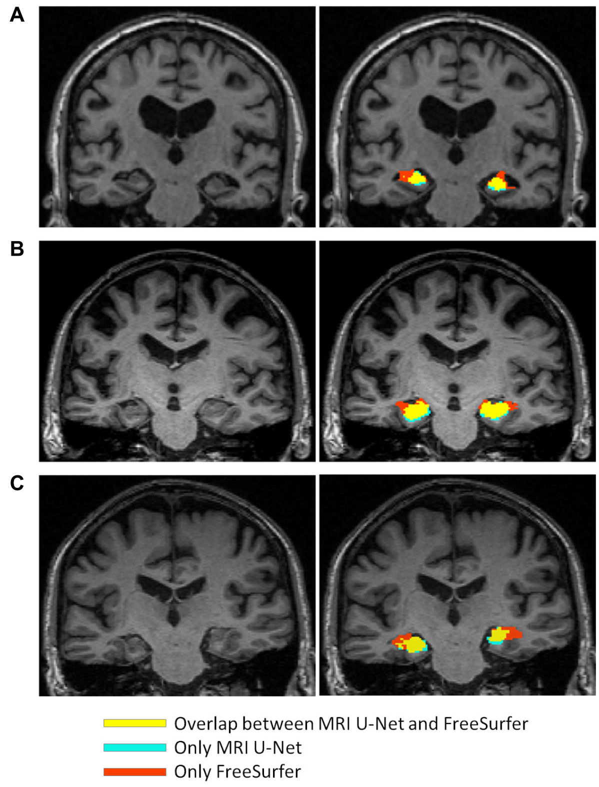


FIGURE 6 | Comparison between the segmentation result obtained from using MRI U-Net and the one from FreeSurfer 6.0 in some examples slices of different subjects from the large ADNI test dataset. The yellow areas indicate the overlap between the two segmentation results, the light blue ones correspond only to the MRI U-Net segmentation, and the red ones only to the FreeSurfer segmentation. All represented subjects belong to the group of 14 cases for which the ratio between the deep-learning based volumes (from all three pipelines) and the volume from the FreeSurfer segmentation had a value that was considered as outlier. In **(A)** and **(B)**, FreeSurfer over-estimates the hippocampal region compared to MRI U-Net that achieves a more reliable estimation. In **(C)**, an unusual shape of the anatomical gyri surrounding the hippocampus is present and both methods result in some segmentation errors, but MRI U-Net achieves a more accurate result compared to the large overestimation obtained by using FreeSurfer.

the second U-net is dealing with a much smaller field of view therefore is less likely to be disturbed by imaging or structure changes outside the core region. On the other hand, compared to Cropped MRI U-Net, the inclusion of shape context layers was also shown to lead to slight, but yet statistically significant, improvements in terms of Dice score and precision. This result supports what has already been observed in the previous section: when a high accuracy in the segmentation is achieved from the simple MRI U-Net implementation, adding shape information does not improve the result; on the other hand, when the MRI U-Net segmentation is more challenging (for example, in this case, due to discrepancies between training and test data), the shape context layers—together with the cropping step—can help to increase the accuracy.

The computation of the evaluation metrics for each independent 2D U-Net also allowed to highlight certain differences between the three approaches that cannot be captured from their global performance. In general, by analyzing the performance of each view independently, the advantage of using Cropped MRI U-Net over MRI U-Net is less noticeable, given its larger differences between views in terms of accuracy, as well as generally higher Hausdorff distance. However, as described above, merging the information from all views together seem to stabilize the result and discard many of the FP that affect the simple 2D-based results. This highlights the importance of integrating information from all views together in order to obtain more reliable segmentation results. This is very consistent with what is suggested for the manual HarP segmentation protocol, i.e., the segmentation must be performed using all views together in order to achieve accurate results. Moreover, similarly to what was observed for the single-cohort analysis, the inclusion of shape context appears again useful to improve the performance not only globally in 3D, but also on a 2D basis. Its performance on all views is indeed superior to the one of both Cropped MRI U-Net and of MRI U-Nets.

The positive contribution of adding the step of shape model fitting is further supported by the comparison with two other types of context information. Indeed, Shape MRI U-Net was found to be the most successful method among all those tested. The difference between Shape MRI U-Net and all other approaches was indeed shown to be statistically significant for most of the evaluation metrics. In particular, as regards Autocontext MRI U-Net, we believe that the network tends to learn mainly from the first U-Net-based segmentation without extracting much more information from the T1 volume. This would explain why there is no real improvement in performance and its accuracy is rather similar to that of MRI U-Net. In the case of Tissue MRI U-Net, instead, we think that automatic tissue types segmentations may tend to fail in some locations. Therefore, this would provide "misleading" information as input to the network, which makes this approach not robust.

The above-discussed observations could be made also when training and test set were switched. Indeed, the average evaluation metrics were quite consistent to those of the first cross-cohort analysis and, also in this case, Shape MRI U-Net showed, overall, the best performance. Only two main differences could be found compared to the previous analysis. First of all, the accuracy of Tissue MRI U-Net got much worse. This could be justified by the fact that, in the AddNeuroMed dataset, the image quality is generally lower, also because of the field strength that is limited to 1.5 T in all subject. This may lead to more imprecise tissue type segmentations used during training, which cause a further degrading of the performance during the test phase on a new dataset. Moreover, for all the other DL-based pipelines, the precision and the recall were, respectively, higher and lower compared to the previous analysis. This result was expected because the training and test sets have been simply switched and therefore possible over-estimations in the first cross-cohort evaluation are likely to result in under-estimations in the second one.

Finally, it should be noted that, despite the increase in performance with Shape MRI U-Net on both cross-cohort

TABLE 5 | Volumetric differences in the hippocampal volume between diagnostic groups in a subset of subjects from the large ADNI test dataset.

Region of interest	Segmentation method	AD (*n* = 93)	MCI (*n* = 267)	HC (*n* = 154)	*p*-value (one-way ANOVA)
Left hippocampus	MRI U-Net	3.54 ± 0.66 cm^3	3.95 ± 0.60 cm^3	4.30 ± 0.54 cm^3	$p < 0.001$
	Shape MRI U-Net	3.67 ± 0.60 cm^3	4.04 ± 0.59 cm^3	4.39 ± 0.51 cm^3	$p < 0.001$
	FreeSurfer 6.0	3.25 ± 0.60 cm^3	3.60 ± 0.63 cm^3	3.99 ± 0.58 cm^3	$p < 0.001$
Right hippocampus	MRI U-Net	3.38 ± 0.66 cm^3	3.81 ± 0.66 cm^3	4.25 ± 0.54 cm^3	$p < 0.001$
	Shape MRI U-Net	3.56 ± 0.61 cm^3	3.93 ± 0.62 cm^3	4.34 ± 0.50 cm^3	$p < 0.001$
	FreeSurfer 6.0	3.16 ± 0.59 cm^3	3.50 ± 0.63 cm^3	3.89 ± 0.54 cm^3	$p < 0.001$

For each subject, the hippocampal volume was multiplied by the ratio between the average ICV and the specific subject's ICV. Results are reported for baseline measurements as mean ± standard deviation for each of the three diagnostic groups of interest, i.e., AD patients, MCI subjects, and healthy controls. Only subjects whose diagnosis did not change within 2 years after the first measurement were selected. The number of subject n in each group is indicated in brackets. For each method, a one-way ANOVA test was conducted for comparing the three diagnostic groups.

TABLE 6 | Prediction power of using the normalized hippocampal volume measurements to classify AD vs. HC, AD vs. MCI, and MCI vs. HC.

Segmentation method	AD vs. HC	AD vs. MCI	MCI vs. HC
MRI U-Net	AUC = 0.85	AUC = 0.68	AUC = 0.67
	Sensitivity = 0.75	Sensitivity = 0.60	Sensitivity = 0.62
	Specificity = 0.82	Specificity = 0.65	Specificity = 0.69
Shape MRI U-Net	AUC = 0.84	AUC = 0.68	AUC = 0.65
	Sensitivity = 0.73	Sensitivity = 0.65	Sensitivity = 0.59
	Specificity = 0.80	Specificity = 0.60	Specificity = 0.66
FreeSurfer 6.0	AUC = 0.82	AUC = 0.68	AUC = 0.64
	Sensitivity = 0.73	Sensitivity = 0.66	Sensitivity = 0.60
	Specificity = 0.73	Specificity = 0.60	Specificity = 0.62

The diagnostic prediction power was analyzed by fitting three different logistic regression model (one for each binary classification case) and computing its AUC, sensitivity, and specificity. The model was fitted to give the probability of a subject to belong to a certain diagnostic group as a function of the hippocampal volume multiplied by the ratio between the average ICV and the specific subject's ICV. Sensitivity and specificity were computed at a threshold of 0.5.

analyses, the segmentation accuracy is still lower than the one obtained using cross-validation on the dataset from the ADNI cohort (presented in section "Single-Cohort Evaluation"). However, this was expected due to both the above-discussed discrepancy between training and test cohort, as well as the inter-rater differences when generating the ground-truth segmentations. The experience of the rater (in terms of familiarity with the segmentation task itself, the given image quality and the specific MRI protocol) can indeed affect the manual delineation of the segmentation masks.

Analysis on the Larger ADNI Dataset

For the last and largest dataset, where ground-truth masks were not available and visually checking the accuracy was not feasible due to the large amount of data, the performance of the networks was checked by comparing the hippocampal volumes with those obtained using FreeSurfer. This approach clearly has limitations, since it cannot give a detailed measure of the accuracy of the method in this new dataset and could not reveal relevant differences between the three proposed methods. However, the high correlation coefficients (presented in **Figure 5**) and similarity metrics (**Supplementary Table S4**) between the proposed methods and FreeSurfer suggest both a valid and consistent performance for all the present methods.

From **Figure 5**, it is possible to observe that FreeSurfer tends to provide, in general, smaller segmentations compared to the DL-based methods. This is in agreement with what was discussed in Section "Cross-Cohort Analysis" when analyzing the performance on the AddNeuroMed dataset. Indeed, in this case, FreeSurfer was shown to have higher average precision and lower average recall.

Furthermore, the number of identified outliers was low in comparison with the size of the dataset, which further supports the consistency of the results across subjects. On the other hand, the visual inspection of some subjects identified as outliers actually revealed a segmentation result from MRI U-Net that did not appear to be less accurate than the one obtained from FreeSurfer, as shown in **Figure 6**. This fact further exposes the limitations of not having a ground-truth mask to validate the performance. On the other hand, it also suggests that the results of the proposed DL-based approaches are promising in comparison with other established methods, especially when dealing with potential clinical cases (since no outliers belonged to the HC group).

When comparing FreeSurfer with the two proposed DL-based methods in terms of Dice score and Hausdorff distance, a rather high consistency could also be observed, especially for Shape MRI U-Net that showed, in average, a higher Dice score. The resulting metrics are also rather consistent with the results obtained when analyzing the performance of FreeSurfer both in the single-cohort and the cross-cohort analyses. This was expected because, as opposed to FreeSurfer, the present U-Net-based methods were all trained on the HarP protocol used for the manual segmentations too.

The availability of pairs of scans acquired from the same subjects at the same time point also allowed us to perform a test–retest analysis. This resulted in a very high CCC (i.e., between 0.977 and 0.989) in the hippocampal volumes between two subsequent scans with both the tested methods, i.e., MRI U-Net and Shape MRI U-Net. While the results obtained in the above-described single- and cross-cohort analyses show the accuracy of the method, these high coefficients in the test–retest investigation demonstrate the reproducibility of the proposed techniques. Moreover, FreeSurfer also resulted in slightly lower

CCCs (between 0.963 and 0.969), showing how the present methods are to some extent more reproducible.

Comparison Between Diagnostic Groups

The performance of the three proposed methods was proven to be satisfactory in all three analyzed diagnostic groups: HC, MCI, and AD patients. By computing the Dice score separately for each group, we found that the segmentation accuracy is quite consistent across the groups.

When the network was initially tested through cross-validation on the dataset of 54 subjects from the ADNI cohort, the AD patients' group was the only one showing a lower performance with respect to the other two. However, this difference was rather small, approximately 1%. The difference between subject groups slightly increased when testing the network on the dataset from the AddNeuroMed cohort, which was different from the cohort of the training data. Indeed, in this case, a little loss in performance was seen already in the MCI subjects, for which the Dice score showed an average decrease of between 1 and 3% compared to HC. In AD patients, the average decrease was between 2 and 4% relative to HC. Observing a slightly better accuracy in the images from HCs was expected. Indeed, MCI subjects and, even more, AD patients are expected to present patterns of hippocampal atrophy, which is strongly related to the severity stage of the disease. These patterns are likely to be quite heterogeneous if compared to the typical hippocampal structure that can be observed on a healthy brain, making the learning of the network more challenging for such diagnostic cases.

Despite the little loss in performance on AD patients, the accuracy of the proposed methods was more satisfactory than the one obtained by applying the automatic segmentation pipeline from FreeSurfer 6.0. FreeSurfer is also affected by a loss in accuracy when segmenting AD patients compared to HC and the magnitude of such loss was always higher than the ones obtained from the presented DL pipelines. These results suggest that the choice of a U-Net-based approach could also be favorable when good segmentation accuracy is needed on brain images from dementia patients. This aspect is particularly important for a medical segmentation tool to be potentially used both in a clinical and a research setting. Indeed, the more accurate the segmentation is, the more reliable the estimations of hippocampal volume and shape will be. Such geometrical features have been shown to be strongly related to the disease progression, and therefore it is crucial to achieve an accurate segmentation also on demented subjects and not only on healthy ones.

The potential of the proposed methods to be used in a clinical framework was also further shown by the comparison between the normalized hippocampal volumes of the three diagnostic groups present in the large ADNI dataset. All present methodologies show significant differences in the distribution of the hippocampal volumes between groups. In particular, the lowest average volume was found in the AD subjects and the highest in the HCs. This suggests that the present methods can capture the differences in volume caused by the atrophy that is typical of the disease progression.

The usefulness of these volumetric differences between groups was further investigated by fitting logistic regression models to predict the diagnosis of a subject. DL-based methods showed a better performance than FreeSurfer 6.0 and the highest AUC (always above 0.80) could be achieved in the classification of AD vs. HC. Similar diagnosis classification tasks have already been investigated in previous literature leading to similar results. Indeed, in a study by Westman et al. (2011b), manual hippocampal segmentations were employed to define multivariate analysis models for diagnosis prediction, obtaining a sensitivity and specificity of, respectively, 87 and 90% for the AD vs. HC classification. Instead, for AD vs. MCI and MCI vs. HC, those evaluation metrics dropped to approximately 70% in all cases. In a later study by Voevodskaya et al. (2014), FreeSurfer 5.1 was used to extract normalized hippocampal volumes from ADNI data and the AUC was computed for three different linear regression models fitted for the same classification tasks. Also in this case, the best result was obtained with AD vs. HC with an AUC of 0.90, while poorer performance was achieved with the other two models. Therefore, our results reflect what has already been observed in literature, i.e., the potential of using accurate hippocampal segmentation methods to improve the diagnosis of AD and its discrimination from healthy cases. Even though there are differences between different studies in their values of AUC, sensitivity, and specificity, it has to be noted that such discrepancies can be due to different factors. First, the number of analyzed subjects and the model definition can highly influence the results, e.g., the model could be affected by overfitting. Moreover, the type and accuracy of the segmentation method being used can also affect the performance. In addition, the patterns of brain atrophy in AD are heterogeneous and it has been estimated that approximately 23% of AD patients are minimally affected by hippocampal atrophy (Poulakis et al., 2018). Therefore, the presence of this type of patients in the dataset can also affect the prediction power of a model based only on hippocampal volume. However, in general, our study is particularly consistent with the others in terms of the difference in performance between the AD vs. HC classification compared to the other two classification tasks. This discrepancy between classifiers, though, will always be expected given the typical patterns of disease progression, since the differences in atrophy between AD and MCI subjects, as well as between MCI and HC, are inevitably smaller compared to the differences between AD patients and healthy subjects.

Computational Time

The present pipelines were proven to be successful not only in terms of segmentation accuracy, but also in terms of computational speed, which varied between approximately 30 and 150 s depending on the architecture being used. Time efficiency is another important aspect to be taken into account in order to use a segmentation tool in a clinical framework as an aid for performing a diagnosis. Therefore, a DL-based solution is promising in the context of potential clinical use. However, it has to be noted that a computationally slower software as FreeSurfer provides, together with the hippocampus, the segmentation masks for many other gray and white matter structures, as opposed to the present study that is focused only on hippocampal segmentation. This implies that the choice of the most efficient segmentation method is strongly dependent on the application of interest and on the level of accuracy that is required from the segmentation result.

Limitations and Future Work

The present work investigates the use of a DL architecture for an image segmentation task that is of particular interest for AD research. Indeed, achieving an accurate hippocampal segmentation is a crucial task for aiding research in the early diagnosis of the disorder. Moreover, precise standards on how to perform a good manual segmentation of the hippocampus are available, making it easier to obtain ground-truth masks to train the network with. However, the number of training data used for the present work was still quite limited. This issue was approached by using data augmentation, but in the future we plan to expand the training dataset by adding more manual segmentations performed by experts. Moreover, it would be useful to obtain manual segmentations of other brain regions, whose geometrical information could be integrated with those from the hippocampus. Therefore, we also aim at extending the study by testing the proposed pipelines on other brain structures that are both of interest for Alzheimer's research and known to be particularly challenging for segmentation, such as the entorhinal cortex. In particular, we want to investigate whether the inclusion of shape information can be even more useful in such a context.

In addition, we would like to change our architecture by using 3D U-Nets instead of the three independent 2D U-Nets. In the present work, an implementation using 2D U-Nets was employed mainly because of the limited 3D training data samples and its advantage over a 3D implementation in terms of memory usage. However, in the future, we would like to test whether the direct use of 3D information could further improve the segmentation accuracy in any of the proposed pipelines.

Moreover, one of the limitations of this study is that the inclusion of shape information encoded in statistical shape models is not entirely new, as already presented in a previous study by Wang and Smedby (2017). In the future, we aim at investigating a wider range of shape descriptors that could possibly further improve the performance of our shape-aware segmentation pipeline. However, besides the different field of application (hippocampal segmentation instead of heart segmentation), the main contribution of this study compared to the one by Wang and Smedby (2017) is the extensive analysis of the performance of Shape MRI U-Net on larger datasets of subjects from different diagnostic groups and cohorts, as well as the comparison with two other types of context-aware architectures. The present work provides a new insight on how the inclusion of *a priori* shape information can be employed in cross-cohort analyses or, more in general, when a testing dataset was not used at the time of training. In the context of hippocampal segmentation, the use of shape information was shown to be indeed more successful than other types of *a priori* information that could be extracted from the given anatomical structures. The integration and comparison with other *a priori* information, as well as the analyses of new cohorts, could be investigated in the future to further confirm the present findings.

Finally, in this study, the MRI scans underwent only a couple of preprocessing stages, i.e., resampling and intensity normalization. A further harmonization of the inputs was later obtained by cropping the images on Cropped MRI U-Net, as well as including the normalized shape models on Shape MRI U-Net. This choice was made to keep the pipeline as simple and quick as possible. However, we would like to investigate whether the addition of a few other preprocessing steps, such as skull stripping, could help improving the performance of MRI U-Net.

CONCLUSION

The present work has proposed an accurate and fast method for automatic segmentation of the hippocampus using U-Net-based DNNs together with statistical shape modeling.

A simpler and quicker U-Net architecture, which simply uses the original MRI scan as input image, achieved already excellent results in a first single-cohort analysis. However, the proposed implementation using shape context was shown to be more successful with data from a new unseen cohort by significantly improving the segmentation accuracy. These results suggest that the inclusion of shape information may make the method more robust in cases where the segmentation task is more challenging.

Our promising results across different diagnostic groups suggest that the proposed method could not only be used as a possible substitute for other existing segmentation tools, but may also have a potential as an aid for studying and diagnosing neurodegenerative disorders.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here: http://adni.loni.usc.edu/.

ETHICS STATEMENT

The data acquisition for the AddNeuroMed study was approved by the Regional Ethics Board Stockholm 2013/694-31. Data were acquired after written informed consent.

AUTHOR CONTRIBUTIONS

CW, ÖS, EW, and IB contributed to the conception and design of the study. OL provided the manual segmentations for the AddNeuroMed dataset. J-SM organized the database and provided the FreeSurfer segmentations. IB and CW developed the segmentation pipeline and shape model. IB, CW, ÖS, EW, and OL worked on the final method evaluation and analysis of the results. All authors contributed to the writing of the manuscript and its revision, as well as approved the submitted version.

FUNDING

This project was financially supported by the Swedish Foundation for Strategic Research (SSF), the Swedish Research Council (VR), the joint research funds of KTH Royal Institute of Technology and Stockholm County Council (HMT), the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Alzheimer Foundation, and the Swedish Brain Foundation. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. AddNeuromed is supported by InnoMed (Innovative Medicines in Europe), an Integrated Project funded by the European Union of the Sixth Framework program priority FP6-2004-LIFESCIHEALTH-5, Life Sciences, Genomics, and Biotechnology for Health.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00015/full#supplementary-material

REFERENCES

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., and Erickson, B. J. (2017). Deep learning for brain MRI segmentation: state of the art and future directions. *J. Digit. Imaging* 30, 449–459. doi: 10.1007/s10278-017-9983-4

Asman, A. J., and Landman, B. A. (2013). Non-local statistical label fusion for multi-atlas segmentation. *Med. Image Anal.* 17, 194–208. doi: 10.1016/j.media.2012.10.002

Boccardi, M., Bocchetta, M., Apostolova, L. G., Barnes, J., Bartzokis, G., Corbetta, G., et al. (2015a). Delphi definition of the EADC-ADNI harmonized protocol for hippocampal segmentation on magnetic resonance. *J. Alzheimers Dement.* 11, 126–138. doi: 10.1016/j.jalz.2014.02.009

Boccardi, M., Bocchetta, M., Morency, F. C., Collins, D. L., Nishikawa, M., Ganzola, R., et al. (2015b). Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimers Dement.* 11, 175–183. doi: 10.1016/j.jalz.2014.12.002

Braak, H., and Braak, E. (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathol.* 82, 239–259. doi: 10.1007/bf00308809

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

Cabezas, M., Oliver, A., Llado, X., Freixenet, J., and Cuadra, M. B. (2011). A review of atlas-based segmentation for magnetic resonance brain images. *Comput. Methods Programs* 104, e158–e177.

Chen, H., Dou, Q., Yu, L., and Heng, P.-A. (2016). Voxresnet: deep voxelwise residual networks for volumetric brain segmentation. *Neuroimage* 170, 446–455. doi: 10.1016/j.neuroimage.2017.04.041

Chen, H., Dou, Q., Yu, L., Qin, J., and Heng, P. A. (2018). VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. *Neuroimage* 170, 446–455. doi: 10.1016/j.neuroimage.2017.04.041

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi: 10.2307/1932409

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

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 33, 341–355.

Fischl, B., Salat, D. H., van der Kouwe, A. J., Makris, N., Segonne, F., Quinn, B. T., et al. (2004). Sequence-independent segmentation of magnetic resonance images. *Neuroimage* 23(Suppl. 1), S69–S84.

Heckemann, R. A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J. V., and Hammers, A. (2010). Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. *Neuroimage* 51, 221–227. doi: 10.1016/j.neuroimage.2010.01.072

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* 15, 850–863. doi: 10.1109/34.232073

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.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Kim, M., Wu, G., Li, W., Wang, L., Son, Y. D., Cho, Z. H., et al. (2013). Automatic hippocampus segmentation of 7.0 Tesla MR images by combining multiple atlases and auto-context models. *Neuroimage* 83, 335–345. doi: 10.1016/j.neuroimage.2013.06.006

Leventon, M. E., Grimson, W. E. L., and Faugeras, O. (2000). "Statistical shape influence in geodesic active contours," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ: IEEE).

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255–268.

Lindberg, O., Walterfang, M., Looi, J. C., Malykhin, N., Ostberg, P., Zandbelt, B., et al. (2012). Hippocampal shape analysis in Alzheimer's disease and frontotemporal lobar degeneration subtypes. *J. Alzheimers Dis.* 30, 355–365. doi: 10.3233/jad-2012-112210

Liu, Y., Paajanen, T., Zhang, Y., Westman, E., Wahlund, L. O., Simmons, A., et al. (2010). Analysis of regional MRI volumes and thicknesses as predictors of conversion from mild cognitive impairment to Alzheimer's disease. *Neurobiol. Aging* 31, 1375–1385. doi: 10.1016/j.neurobiolaging.2010.01.022

Lovestone, S., Francis, P., Kloszewska, I., Mecocci, P., Simmons, A., Soininen, H., et al. (2009). AddNeuroMed–the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Ann. N. Y. Acad. Sci.* 1180, 36–46. doi: 10.1111/j.1749-6632.2009.05064.x

Ma, J., Lin, F., Wesarg, S., and Erdt, M. (2018). "A novel bayesian model incorporating deep neural network and statistical shape model for pancreas segmentation," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, (Cham: Springer International Publishing).

Mahbod, A., Chowdhury, M., Smedby, Ö, and Wang, C. (2018). Automatic brain segmentation using artificial neural networks with shape context. *Pattern Recogn. Lett.* 101, 74–79. doi: 10.1016/j.patrec.2017.11.016

Milletari, F., Ahmadi, S. A., Kroll, C., Plate, A., Rozanski, V., Maiostre, J., et al. (2017). Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. *Comput. Vis. Image Understand.* 164, 92–102. doi: 10.1016/j.cviu.2017.04.002

Mirikharaji, Z., Izadi, S., Kawahara, J., and Hamarneh, G. (2018). "Deep auto-context fully convolutional neural network for skin lesion segmentation," in *Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, (Piscataway, NJ: IEEE).

Muehlboeck, J. S., Westman, E., and Simmons, A. (2013). TheHiveDB image data management and analysis framework. *Front. Neuroinform.* 7:49. doi: 10.3389/fninf.2013.00049

Oliva, A., and Torralba, A. (2007). The role of context in object recognition. *Trends Cogn. Sci.* 11, 520–527. doi: 10.1016/j.tics.2007.09.009

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., et al. (2016). Brain atrophy in Alzheimer's Disease and aging. *Ageing Res. Rev.* 30, 25–48. doi: 10.1016/j.arr.2016.01.002

Pipitone, J., Park, M. T. M., Winterburn, J., Lett, T. A., Lerch, J. P., Pruessner, J. C., et al. (2014). Multi-atlas segmentation of the whole hippocampus and subfields using multiple automatically generated templates. *Neuroimage* 101, 494–512. doi: 10.1016/j.neuroimage.2014.04.054

Poulakis, K., Pereira, J. B., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., et al. (2018). Heterogeneous patterns of brain atrophy in Alzheimer's disease. *Neurobiol. Aging* 65, 98–108.

Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P., and Vaidya, V. (2017). "Learning and incorporating shape models for semantic segmentation," in *Proceedings of the Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, (Cham: Springer International Publishing).

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, (Cham: Springer International Publishing).

Scheltens, P., Fox, N., Barkhof, F., and De Carli, C. (2002). Structural magnetic resonance imaging in the practical assessment of dementia: beyond exclusion. *Lancet Neurol.* 1, 13–21. doi: 10.1016/s1474-4422(02)00002-9

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 640–651. doi: 10.1109/TPAMI.2016.2572683

Simmons, A., Westman, E., Muehlboeck, J. S., Mecocci, P., Vellas, B., Tsolaki, M., et al. (2009). MRI measures of Alzheimer's disease and the addneuromed study. *Ann. N. Y. Acad. Sci.* 1180, 47–55. doi: 10.1111/j.1749-6632.2009.05063.x

Simmons, A., Westman, E., Muehlboeck, S., Mecocci, P., Vellas, B., Tsolaki, M., et al. (2011). The AddNeuroMed framework for multi-centre MRI assessment of Alzheimer's disease: experience from the first 24 months. *Int. J. Geriatr. Psychiatry* 26, 75–82. doi: 10.1002/gps.2491

Tabatabaei-Jafari, H., Shaw, M. E., and Cherbuin, N. (2015). Cerebral atrophy in mild cognitive impairment: a systematic review with meta-analysis. *Alzheimers Dement.* 1, 487–504. doi: 10.1016/j.dadm.2015.11.002

Taha, A. A., and Hanbury, A. (2015). An efficient algorithm for calculating the exact hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 2153–2163. doi: 10.1109/TPAMI.2015.2408351

Tang, M., Valipour, S., Zhang, Z., Cobzas, D., and Jagersand, M. (2017). *A Deep Level Set Method for Image Segmentation. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing.

Thyreau, B., Sato, K., Fukuda, H., and Taki, Y. (2018). Segmentation of the hippocampus by transferring algorithmic knowledge for large cohort processing. *Med. Image Anal.* 43, 214–228. doi: 10.1016/j.media.2017.11.004

Tu, Z., and Bai, X. (2010). Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 1744–1757. doi: 10.1109/TPAMI.2009.186

Vinters, H. V. (2015). Emerging concepts in Alzheimer's disease. *Annu. Rev. Pathol.* 10, 291–319. doi: 10.1146/annurev-pathol-020712-163927

Voevodskaya, O., Simmons, A., Nordenskjöld, R., Kullberg, J., Ahlström, H., Lind, L., et al. (2014). The effects of intracranial volume adjustment approaches on multiple regional MRI volumes in healthy aging and Alzheimer's disease. *Front. Aging Neurosci.* 6:264. doi: 10.3389/fnagi.2014.00264

Wang, C., and Smedby, Ö (2017). *Automatic Whole Heart Segmentation Using Deep Learning and Shape Context. International Workshop on Statistical Atlases and Computational Models of the Heart*. Cham: Springer.

Wang, H., and Yushkevich, P. A. (2013). Multi-atlas segmentation with joint label fusion and corrective learning-an open source implementation. *Front. Neuroinform.* 7:27. doi: 10.3389/fninf.2013.00027

Westman, E., Simmons, A., Muehlboeck, J. S., Mecocci, P., Vellas, B., Tsolaki, M., et al. (2011a). AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. *Neuroimage* 58, 818–828. doi: 10.1016/j.neuroimage.2011.06.065

Westman, E., Simmons, A., Zhang, Y., Muehlboeck, J. S., Tunnard, C., Liu, Y., et al. (2011b). Multivariate analysis of MRI data for Alzheimer's disease, mild cognitive impairment and healthy controls. *Neuroimage* 54, 1178–1187. doi: 10.1016/j.neuroimage.2010.08.044

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

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 Neuroscience

METHODS
published: 20 March 2020
doi: 10.3389/fnins.2020.00179

DeepBrainSeg: Automated Brain Region Segmentation for Micro-Optical Images With a Convolutional Neural Network

Chaozhen Tan[1,2], *Yue Guan*[1,2], *Zhao Feng*[1,2], *Hong Ni*[1,2], *Zoutao Zhang*[1,2], *Zhiguang Wang*[1,2], *Xiangning Li*[1,2,3], *Jing Yuan*[1,2,3], *Hui Gong*[1,2,3], *Qingming Luo*[1,2] *and Anan Li*[1,2,3]*

[1] Britton Chance Center for Biomedical Photonics, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China, [2] MoE Key Laboratory for Biomedical Photonics, School of Engineering Sciences, Huazhong University of Science and Technology, Wuhan, China, [3] HUST-Suzhou Institute for Brainsmatics, JITRI Institute for Brainsmatics, Suzhou, China

OPEN ACCESS

Edited by:
Amir Shmuel,
McGill University, Canada

Reviewed by:
Ulas Bagci,
University of Central Florida, United States
Fatemeh Zabihollahy,
Carleton University, Canada

Correspondence:
Anan Li
aali@mail.hust.edu.cn

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *24 July 2019*
Accepted: *18 February 2020*
Published: *20 March 2020*

Citation:
Tan C, Guan Y, Feng Z, Ni H, Zhang Z, Wang Z, Li X, Yuan J, Gong H, Luo Q and Li A (2020) DeepBrainSeg: Automated Brain Region Segmentation for Micro-Optical Images With a Convolutional Neural Network. Front. Neurosci. 14:179. doi: 10.3389/fnins.2020.00179

The segmentation of brain region contours in three dimensions is critical for the analysis of different brain structures, and advanced approaches are emerging continuously within the field of neurosciences. With the development of high-resolution micro-optical imaging, whole-brain images can be acquired at the cellular level. However, brain regions in microscopic images are aggregated by discrete neurons with blurry boundaries, the complex and variable features of brain regions make it challenging to accurately segment brain regions. Manual segmentation is a reliable method, but is unrealistic to apply on a large scale. Here, we propose an automated brain region segmentation framework, DeepBrainSeg, which is inspired by the principle of manual segmentation. DeepBrainSeg incorporates three feature levels to learn local and contextual features in different receptive fields through a dual-pathway convolutional neural network (CNN), and to provide global features of localization by image registration and domain-condition constraints. Validated on biological datasets, DeepBrainSeg can not only effectively segment brain-wide regions with high accuracy (Dice ratio > 0.9), but can also be applied to various types of datasets and to datasets with noises. It has the potential to automatically locate information in the brain space on the large scale.

Keywords: automated segmentation, brain regions, convolutional neural networks, image registration, domain-condition constraints, micro-optical images

INTRODUCTION

Complex structures in the brain have the specificity for brain regions, which correspond to varying brain functions. The maturation of techniques for high-resolution micro-optical imaging (Li et al., 2010; Ragan et al., 2012; Gong et al., 2016) has allowed comprehensive measurements of the distributions of fine structures in three-dimensional (3D) brain space. This has led to better understanding of brain structures, such as whole-brain neuron projections (Economo et al., 2016; Li et al., 2018), cellular and vascular distributions (Peng et al., 2017; Xiong et al., 2017). Such analyses require 3D brain region contours as boundary preconditions. However, unlike magnetic resonance images (MRIs), brain regions in microscopic images are aggregated by discrete neurons, resulting in blurry boundaries between regions (Gahr, 1997). Identifying the boundaries requires to combine a number of features, including cellular staining, morphology, and distribution. Moreover, due to individual differences and imaging processes, these complex features are variable, making it challenging to accurately segment brain regions. The manual segmentation of brain region contours (Dong, 2008) by anatomists is considered to be a reliable method, but is unrealistic to apply on a large scale for high-resolution images. Therefore, neuroscientists urgently require an automated and accurate method that can segment brain regions at the cellular level.

Image segmentation has been studied extensively for brain sciences. Classic segmentation methods (Clarke et al., 1995; Balafar et al., 2010; Nanthagopal and Sukanesh, 2013) based on hand-crafted features have been used for a long time and primarily utilize the differences between features, such as intensity and texture. For example, Feng et al. (2017) used a 3D Otsu method with intensity features to segment MRI brain structures. However, the features of brain regions for micro-optical images are complex, and vary between different individuals and imaging devices, rendering the hand-crafted features approach inappropriate for micro-optical brain images.

Deep learning (LeCun et al., 2015; Schmidhuber, 2015; Shen et al., 2017) for image segmentation is another rapidly developing field. Methods based on convolutional neural networks (CNNs) (Krizhevsky et al., 2012; Rawat and Wang, 2017) can build complex deep-level features based on simple low-level features, making them competitive against classic shallow hand-crafted features approaches. One approach for image segmentation which uses CNNs has an end-to-end form with full convolutions (Long et al., 2015; Milletari et al., 2016; Badrinarayanan et al., 2017; Chen et al., 2017; Jégou et al., 2017; Yu et al., 2017; Chen et al., 2018); i.e., the output of the network is the result of pixel-by-pixel segmentation. For instance, U-net (Ronneberger et al., 2015), consisting of groups of convolutional and deconvolutional layers and skip links, is widely applied in medical image segmentation. Whereas, due to pooling layers, the end-to-end approach may adversely affect the image resolution and therefore result in loss of details (Litjens et al., 2017). Moreover, since a whole image constitutes one sample, many hours of labor are required to label enough samples for training.

Another CNN approach, the patch-based method (Lai, 2015; Pereira et al., 2016), is able to handle the details and label samples to an acceptable level. This approach classifies each pixel in the image individually by presenting it with patches extracted around that particular pixel (Litjens et al., 2017). For example, Ciresan et al. (2012) used a patch-based CNN to segment medical images; furthermore, multi-scale CNNs (de Brebisson and Montana, 2015; Moeskops et al., 2016) were adopted to achieve a higher accuracy for MR brain images with different receptive fields (Luo et al., 2016). However, the patch-based approach has the limitations of low efficiency and lack of global information.

Neuroanatomical studies benefit from the ability to obtain high-resolution micro-optical images, which allows fine division of the brain into thousands of regions (Kuan et al., 2015). The steps for manual segmentation of brain regions by anatomists consist in locating the structure at the macroscale, identifying the shape and neighboring differences at the mesoscale, and segmenting accurate boundaries at the microscale. Correspondingly, the automated segmentation also requires multi-level features: global, contextual, and local. While CNN methods can learn local and contextual features, they have difficulty utilizing global location features from the whole-brain range at high resolution, resulting in over-segmentation for other regions with similar local features. To locate brain structures, Iqbal et al. (2019) segmented and classified the mouse brain into eight regions using Mask r-cnn (He et al., 2017), while the detected box has excessive redundancies for the region with complex shape. Chen et al. (2019) combined a patch-based CNN and registration to segment the murine brainstem, whereas the accuracy of segmentation is easily affected by the effect of registration. In other words, current automated methods are not capable of utilizing on global, contextual, and local information to accurately segment brain regions for micro-optical images.

We propose a framework inspired by the principle of manual segmentation, DeepBrainSeg, which automatically locates and segments brain regions incorporating three level features: local, contextual, and global. We design a dual-pathway network with two-scale patches to acquire local and contextual features in different receptive fields, and combine image registration and domain-condition constraints for initial and tracking localization. We segmented several brain-wide regions and quantitatively evaluated the segmentation effect: which shows a high accuracy (Dice ratio $>$ 0.9). DeepBrainSeg achieves more accurate results than U-net, V-net, FC-DensNet, and Segnet. It is also suitable for datasets with noises and can be used for various types of datasets. In addition, DeepBrainSeg demonstrates high computational efficiency on different platforms.

MATERIALS AND METHODS

Biological Datasets

In this study, we used 14 mouse brain datasets from four different imaging systems. Ten datasets are Thy1-GFP M-line transgenic mice whose whole brains are imaged using a dual color fluorescence microscope [Brain-wide Precision Imaging system (BPS)] (Gong et al., 2016). The other four datasets are a Nissl-stained C57BL/6 adult mouse imaged using a Micro-Optical Sectioning Tomography (MOST) system (Li et al., 2010), a C57BL/6 mouse with autofluorescent signal imaged with a serial two-photon (STP) system (Ragan et al., 2012), a C57BL/6 adult mouse imaged with MR image model T2* (Johnson et al., 2010), and the Allen mouse common coordinate framework (Allen CCF v3 brain atlas) containing an 3D average brain image and a labeled brain region space. We got the STP dataset from "http://www.swc.ucl.ac.uk/aMAP," the MR dataset from "civmvoxport.vm.duke.edu," and the Allen CCF from "https://atlas.brain-map.org." The pixel resolution of the MR dataset is 21.5 μm isotropic; others are all sampled to 10 μm isotropic.

The Framework of DeepBrainSeg

DeepBrainSeg consists of three parts (**Figure 1**): network training, initial localization, and predicting with tracking localization. First, we obtain images and labels by manually delineating the boundaries of brain regions, screen and augment the samples to generate the training set, and train the designed dual-pathway CNN (**Figure 1A**). Then, for the new unlabeled image, we perform a 3D registration with Allen CCF, and map the label from the Allen CCF to the unlabeled image, select one two-dimensional (2D) label slice and dilate it as the initial localization of the brain region (**Figure 1B**). Finally, the located 2D image is used for predicting by the trained CNN, and the segmentation result is dilated as the domain-condition constraint to locate the adjacent images. Tracking localization and prediction are performed alternately until the complete 3D segmentation results are obtained (**Figure 1C**).

Label and Sample Extraction

The main datasets used for verification in this study are 10 datasets from BPS. Five brain regions with visible differences in the surrounding areas were selected for training and predicting: main olfactory bulb mitral layer (MOBmi), pyramidal and granular layers of the hippocampus (HIP-pg), the granular layer of cerebellar cortex (CBX-gl), outline, and facial nerve (VIIn). For each brain region, 100 coronal planes from five datasets were selected at intervals as the training and predicting images. Subsequently, using the Amira (version 6.1.1; FEI, Mérignac Cedex, France) tool, three experienced technicians generated the "labels" by manually demarcating the boundaries of the brain region on training and predicting images, to be used as the ground truth (**Figure 1A**.1).

For the dual-pathway CNN, a sample is presented as images with two different sizes around the particular pixel, and the

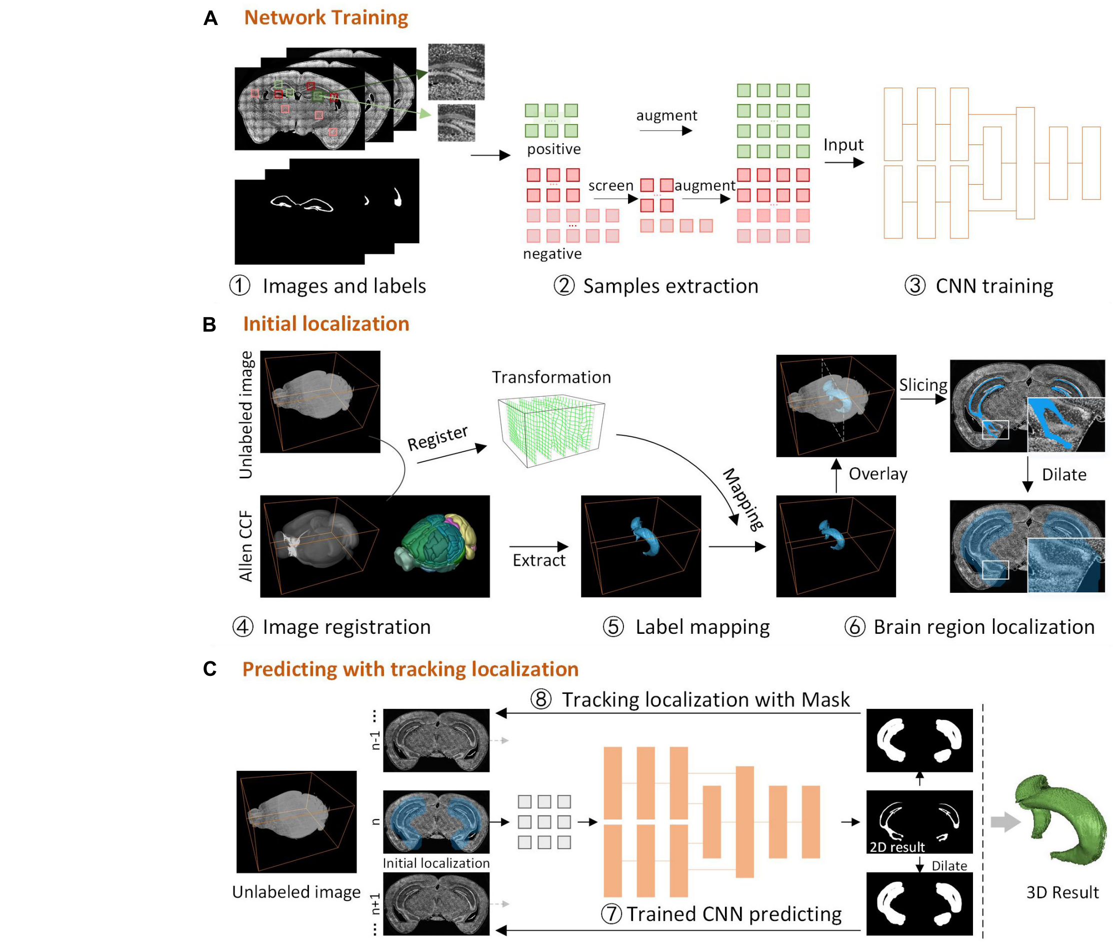


FIGURE 1 | The framework for DeepBrainSeg. **(A)** Network training, the acquisition of images and labels, samples extraction, building and training the CNN. **(B)** Initial localization, image registration for the unlabeled image and Allen CCF, mapping the label to the image, and initial localization the brain region. **(C)** Predicting with tracking localization, predicting the initial image, dilating the 2D result as the localization of adjacent images, alternating prediction and localization to obtain a 3D result.

value of the pixel in the label is the classification. There are two common problems in the sample extraction: there is much redundancy between adjacent patches, and the number of patches where the center pixel is within the brain regions (the positive samples) is much smaller than in other regions (the negative samples). To solve these problems, we customized the sample extraction scheme according to the characteristics of the brain regions (**Figure 1A**.2). First, we extract samples at intervals on coronal images to avoid excessive repetitive information. Then the data are screened and augmented (Krizhevsky et al., 2012) to maintain the equilibrium of positive and negative samples. The augmentation extends the intensity range in the data to improve the ability of the model for generalization. The process is as follows: randomly remove 90% of negative samples containing no pixel in the brain regions; randomly remove *x*% negative samples containing parts of pixels in the brain regions; augment the rest of samples by increasing and decreasing the intensity by 20%. The equilibrium of positive and negative samples is as follows:

$$10\%N_1 + 3 \cdot x\%N_2 = 3N_3 \tag{1}$$

where N_1 and N_2 are the number of negative samples containing no pixel and parts of pixels in the brain regions, respectively, and N_3 is the number of positive samples. Finally, we extracted hundreds of thousands of training samples for each brain region, of which 80% were used as the train set and 20% as the validation set.

Dual-Pathway CNN Training

In order to acquire the local and contextual features from different receptive fields, we designed a dual-pathway CNN with two-scale patches to segment brain regions. The smaller patches mainly provide local features while the larger patches provide contextual features. As shown in **Figure 2**, the network first consists of two same-pathway structures with three hidden layers. The first two hidden layers consist of a convolutional layer, an active layer, a local response normalization (LRN), and a pooling layer. The convolution kernel is 5 × 5, the stride is 1 × 1, the activation layer uses rectified linear units (ReLUs), the pooling layer uses 3 × 3 max-pooling, and the stride is 2 × 2. The third hidden layer consists of a convolutional layer and an active layer. The two-pathway network results in 128 5 × 5 feature maps. Subsequently, the feature maps are cascaded and connected by a 5 × 5 convolutional layer and a ReLU to acquire 512 1 × 1 feature maps. Then, the feature maps from the third and the fourth layer hidden layers are input to the corresponding fully connected layers. All the feature maps are concatenated and input to a fully connected layer. Following this, ReLU is applied, and dropout is used to prevent overfitting. Finally, the SoftmaxWithLoss classifier is used to handle the feature maps. The softmax function is defined as follows:

$$\sigma_i(z) = \frac{\exp(z_i)}{\sum_{j=1}^{m}\exp(z_j)}, i = 1, \ldots, m \tag{2}$$

The multinomial logistic loss function is defined as

$$\ell(y, o) = -\log(o_y) \tag{3}$$

Finally, the combined softmax and loss function are expressed as

$$\widetilde{\ell}(y, z) = -\log(\frac{e^{z_y}}{\sum_{j=1}^{m} e^{z_j}}) = \log(\sum_{j=1}^{m} e^{z_j}) - z_y \tag{4}$$

The network training was implemented through Caffe (Jia et al., 2014) to obtain five models of the corresponding brain regions. During the training process, the batch size is 200, and the maximum number of iterations is 50,000 with 100 epochs. The learning rate is initialized to 0.01, and the iterative decay algorithm by step is applied every 10,000 iterations. The momentum and weight decay are 0.9 and 0.0005, respectively. The training is executed on the GPU to improve the efficiency.

Initial Localization by Image Registration

It is necessary to locate brain regions before predicting the segmentation result to avoid over-segmentation and to improve efficiency. Brain atlas is commonly used as a reference for brain region recognition. Here, we use Allen CCF to locate the brain region by mapping the segmented labels to new images. Allen CCF consists of a 3D average brain image and a corresponding labeled brain region space. First, we register the unlabeled image and the average brain in 3D to obtain the transformation (**Figure 1B**.4). Then, the label for corresponding brain region from Allen CCF is extracted, and the label is mapped to the new image with the transformation (**Figure 1B**.5), which enables general localization of brain regions. However, due to differences in biological samples and imaging mode, it can be difficult to guarantee an accurate match between the mapped label and brain region, especially where brain regions appear and disappear. Instead of locating the whole 3D brain region, we select a 2D label from the middle slice of the 3D label as the initial localization and then perform a dilation of the label to eliminate registration errors, which ensures that all pixels within the brain region are included in the dilated label (Mask-init) (**Figure 1B**.6).

For image registration, a multi-resolution pyramid strategy is used for acceleration. Each hierarchy contains both linear and non-linear registration, and aims to maximize mutual information between the unlabeled image and the average brain. Symmetric diffeomorphic normalization (Avants et al., 2008), a widely used method, is conducted as the non-linear transformation model. Its energy function is defined as

$$E_{sym}(I, J) = \inf_{\phi_1}\inf_{\phi_2}\int_{t=0}^{0.5}\{||v_1(x, t)||_L^2 + ||v_2(x, t)||_L^2\}dt + \int_{\Omega}|I(\phi_1(0.5) - J(\phi_2(0.5)|^2 d\Omega \tag{5}$$

where v_1 and v_2 are the velocity field in opposite directions and $\emptyset_1$and $\emptyset_2$ are the diffeomorphism field in opposite directions.

Simultaneous Tracking Localization and Prediction

The 3D brain region can be regarded as changes of the 2D brain region slice in the spatial domain. High axial resolution imaging

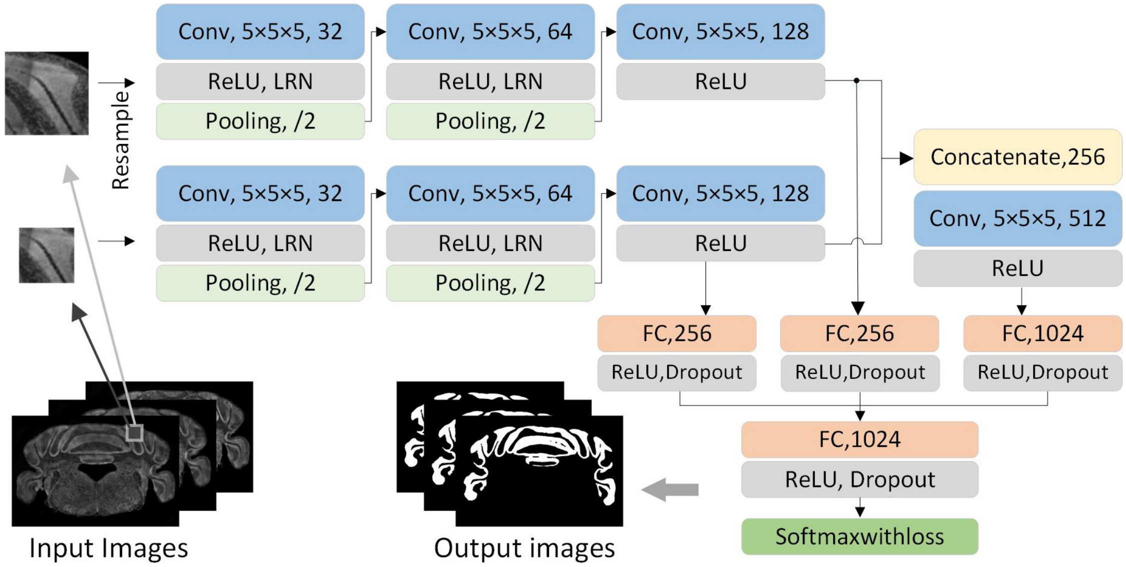


FIGURE 2 | The architecture of dual-pathway CNN. The network consists of dual pathways that take the smaller and larger patch as input, respectively. Each pathway has three hidden layer which have the main components of a convolutional layer, a ReLU layer, an LRN layer, and a pooling layer. The dual-pathway feature maps form the input to a full connection and a convolution layer. All are concatenated after a full connection, and the SoftmaxWithLoss classifier is applied at the end.

makes adjacent 2D brain region slice change less, making it possible to track the 2D brain region in a similar way to target tracking on a video in the time domain. Based on this idea, we proposed a strategy to locate the brain region during the prediction. For 3D brain region segmentation, initial localization by image registration is performed as the first predicting image with Mask-init, patches in the Mask-init are extracted from the 2D image as the input of trained CNN, and the 2D segmentation result is obtained through network predicting (**Figure 1C**.7). Subsequently, we dilate the 2D result as the domain-condition constraint (Mask-track) of the adjacent images (**Figure 1C**.8). For adjacent images, the network predicts patches in the Mask-track to get the 2D result. Finally, alternate tracking localization and prediction are performed for the rest of the corresponding 2D images to obtain a 3D segmentation result, and postprocessing operations including hole filling, connected component analysis, and 3D smoothing are conducted. **Figure 3** demonstrates the segmentation effect with and without Mask-track, localization can avoid over-segmentation of similar local features.

In addition, only the pixels in the Mask-track require predicting, which greatly improves the efficiency. Moreover, based on the connected domain characteristics, brain regions are predicted at one pixel interval to reduce computation by three quarters. These optimizations make it efficient for high-resolution images.

Quantitative Evaluation

To assess the accuracy of our method, we used three parameters to evaluate the segmentation effect: Dice (Dice, 1945), Precision, and Recall. The corresponding formulae are as follows:

$$\text{Dice}\ (I,\ J) = \frac{2 \times |I \cap J|}{|I| + |J|} \tag{6}$$

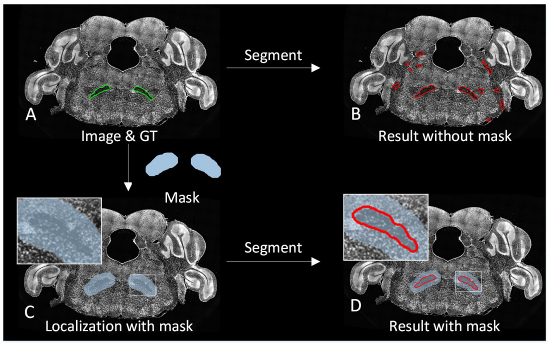


FIGURE 3 | Comparison of the segmentation effect with and without localization. **(A)** A superposition of the original image and the manually segmented lines. **(B)** The segmentation result without localization. **(C)** A superposition of the original image and Mask. **(D)** The segmentation result with localization.

$$\text{Precision}\ (I,\ J) = \frac{|I \cap J|}{|I|} \tag{7}$$

$$\text{Recall}\ (I,\ J) = \frac{|I \cap J|}{|J|} \tag{8}$$

where I and J represent automated and manual binarized segmentation images, | I| and | J| denote the numbers of pixels in brain regions, and | I∩J| denotes the intersection of | I| and | J| for the pixels in brain regions.

In addition, we also quantitatively assess the effect of localization by Precision and Recall, where I represents automatically located binarized images (Mask-init and Mask-track).

Testing Environments

We tested our method on two different computing platforms: A graphical workstation equipped with a NVIDIA M6000 GPU card, 20 CPU cores (Intel Xeon E5-2687w × 2), and 128 GB of RAM. A GPU server equipped with four NVIDIA V100 GPU cards, 12 CPU cores (Intel Xeon xeon-6126w × 2), and 192 GB of RAM.

RESULTS

Determination of Dilation Sizes and Two-Scale Path Size

Here, we experimentally determined the optimal values for parameters by investigating different dilation sizes of Mask-init, Mask-track, and two-scale patch size. Since CBX-gl can be wide-ranging in terms of sizes, we used it as a representative for experiment using 10 BPS datasets. For testing of dilation size, Precision indicates the redundancy of localization range for the ground truth, and Recall indicates the accuracy of the localization. To ensure subsequent segmentation accuracy, Recall must be very close to 1. We therefore assessed sizes of 10, 20, 30, 40, and 50 pixels. As shown in **Table 1**, Recall ratio for Mask-init increases as the dilation size increases, it achieves the highest of 0.999 at 50 pixels, and Recall ratio for Mask-track reaches 0.999 at 20 pixels. Therefore, we determine the optimal dilation sizes for Mask-init and Mask-track with 50 pixel and 20 pixels, respectively.

For two-scale path size, the receptive field will increase as patch size increases with the amount of information in a wide area. This improves classification accuracy, but also reduces positioning accuracy. To obtain optimal patch sizes, two groups of tests were conducted by first determining the smaller size patches and then the larger size using Dice ratio. First, five patch sizes (23, 35, 51, 75, and 91) were chosen to build single-scale networks. As the patch size increases, the wider receptive field improves classification accuracy, decreasing the numbers of outlier, and Dice ratio gradually increases (**Figure 4A**), reaching a peak at 51 pixels^2 (Dice ratio = 0.944), after which it declines. We therefore selected 51 pixels^2 as the optimal parameter for the smaller size. Then, the selected smaller patch size is multiplied by 1.5, 2, or 3 times to produce the larger patch sizes. **Figure 4B** shows the Dice ratio for different multiples. The highest value is obtained with a multiplication factor of 1.5 (Dice ratio = 0.952), after which it declines as the reduction of positioning accuracy has a major impact. Meanwhile, Dice ratio also reveals that the accuracy of two-scale is higher than the single-scale. We ultimately obtained the optimal patch sizes of 51 × 51 and 77 × 77 pixels^2 for brain segmentation.

Segmentation for Brain-Wide Regions

In the field of neuroscience, the analysis of brain space and information commonly requires the segmentation of multiple brain regions which are distributed throughout the brain. Here, we selected five brain regions from ten BPS datasets for segmentation (see section "Materials and Methods" for specific training and prediction procedures). Using the trained models for each brain region, we performed localization and prediction for 50 corresponding images from five datasets. One dataset is used to illustrate the effects of localization and segmentation, by showing the overlapping of the original

TABLE 1 | Performance of Recall ratio for Mask-init and Mask-track with different dilation size (bold values are the optimal).

Dilate size	10 pixels	20 pixels	30 pixels	40 pixels	50 pixels
Mask-init	0.807 ± 0.063	0.936 ± 0.032	0.975 ± 0.017	0.994 ± 0.005	**0.999 ± 0.001**
Mask-track	0.997 ± 0.008	**0.999 ± 0.005**	0.999 ± 0.004	1.000 ± 0.002	

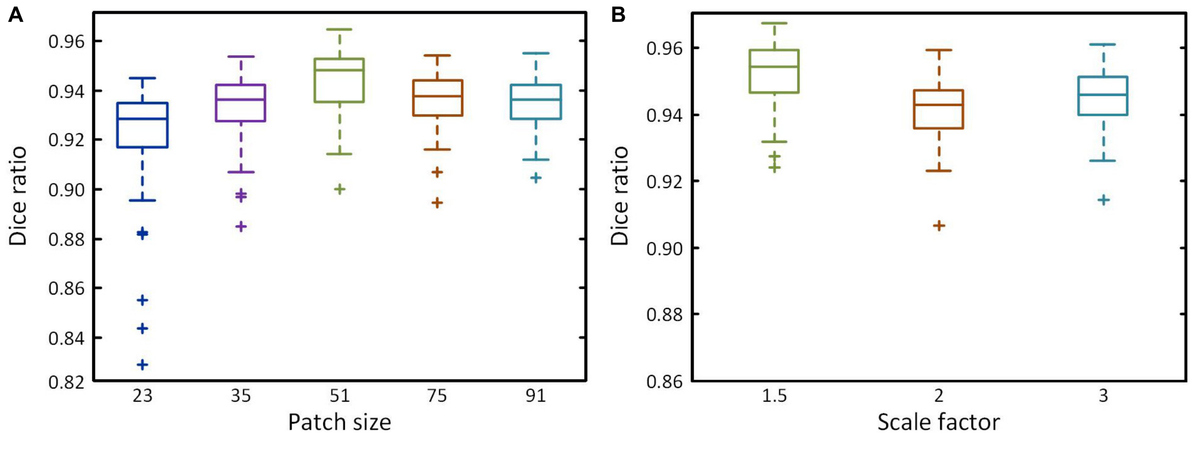


FIGURE 4 | Performance of different patch size. **(A)** Box plots showing the Dice ratio for five different patch sizes at a single scale. **(B)** Box plots showing the Dice ratio for the larger patch at different multiples of the smaller patch size.

images, the located Mask, and the segmented lines from the binarized results (**Figure 5**). **Figure 5A** reveals the overall effects for the five brain regions (MOBmi, HIP-pg, CBX-gl, VIIn, and outline). Although there are differences in the characteristics among each brain region, DeepBrainSeg displays good localization and segmentation effects on all of these regions. **Figures 5B–E** show enlarged images of the white boxes in **Figure 5A**. The segmented lines are close to the real boundaries in the detail images. In particular, HIP-pg and CBX-gl, which have complicated shapes, also maintain fine effects. **Figure 5F** shows a 3D reconstruction of the segmentation results, which demonstrates the integrity and continuity of our approach in 3D space.

We also quantitatively evaluated the performance of localization and segmentation for these 50 images from five brain regions. **Figures 6A,B** show Recall and Precision (Redundancy) ratio for localization. Recall of all brain regions is very close to 1, indicating that almost all pixels of brain regions are included in the Mask, and Redundancy is between 0.14 and 0.83 for different regions. **Figures 6C–E** demonstrate box plots of Dice, Precision, and Recall for the segmentation effect. All three parameters exceed 0.95 for MOBmi, CBX-gl, and outline, and 0.92 for the complex HIP-pg structure. Although subtle deviations in the automated segmentation will affect the parameters for small brain regions, the parameters are consistently above 0.85 for VIIn. Detailed performance statistics showing means and standard deviations are provided in **Table 2**.

Segmentation for Datasets With Noises

For long-term continuous micro-optical imaging, it is easy to generate noises such as stripes and darkened corners through

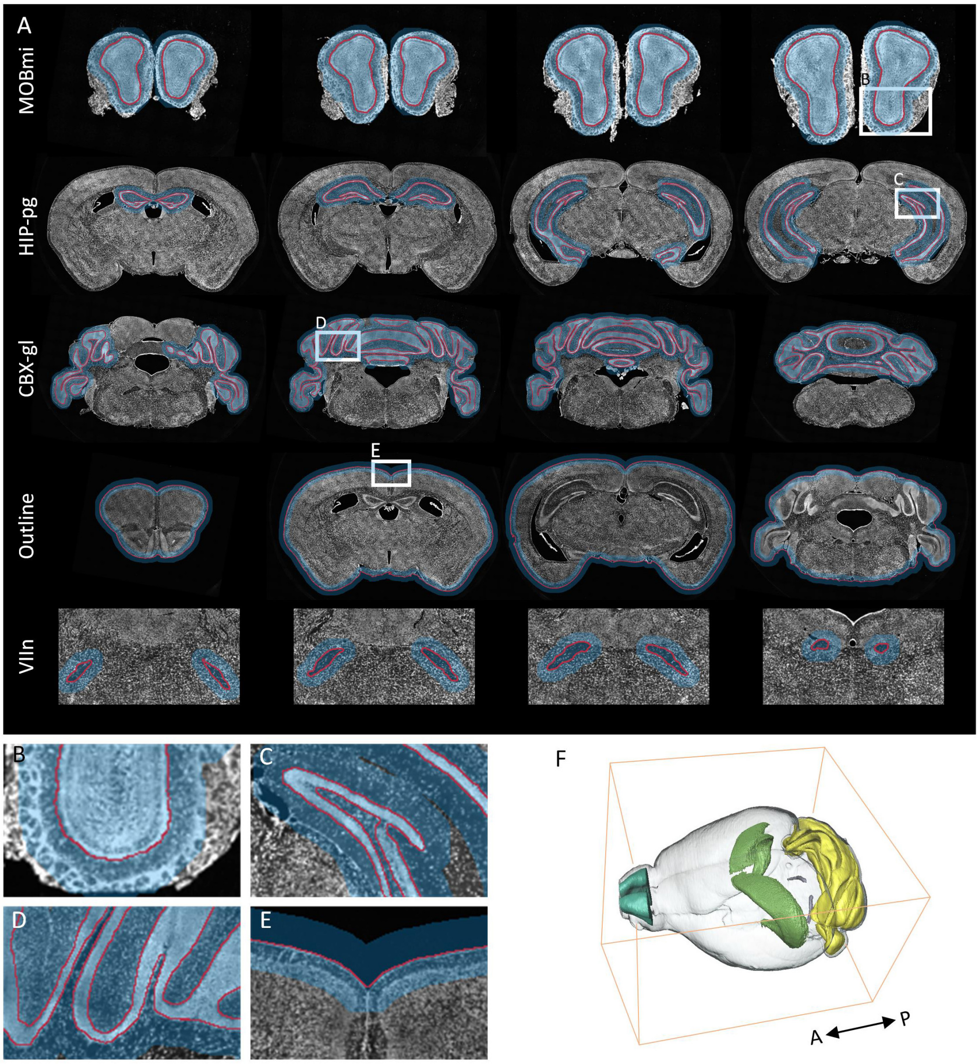


FIGURE 5 | Segmentation effects for brain-wide regions. **(A)** The segmentation effects for five brain regions. From top to bottom: MOBmi, HIP-pg, CBX-gl, VIIn, and outline, each of which are shown as the superposition of four typical coronal images, localization masks, and segmented lines. **(B–E)** Enlarged views of the white boxes in **A**. **(F)** A 3D reconstruction of the segmentation results of the five brain regions.

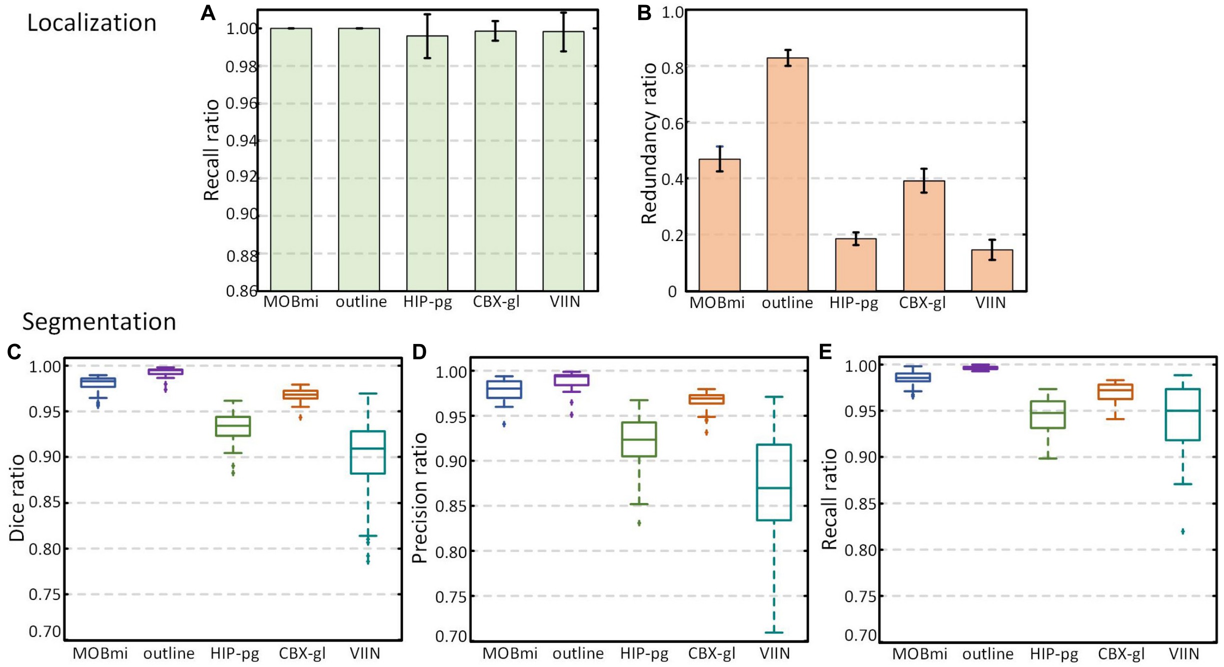


FIGURE 6 | Performance of DeepBrainSeg for brain-wide regions. **(A,B)** Recall and Redundancy of localization effect. **(C–E)** Box plots showing Dice, Precision, and Recall (from left to right) of segmentation effect.

TABLE 2 | Performance of DeepBrainSeg for brain-wide regions.

		MOBmi	Outline	HIP	CB	VIIn
Localization	Recall	1.000 ± 0.000	1.000 ± 0.000	0.996 ± 0.012	0.999 ± 0.005	0.998 ± 0.010
	Redundancy	0.468 ± 0.044	0.828 ± 0.028	0.186 ± 0.023	0.391 ± 0.042	0.146 ± 0.036
Segmentation	Dice	0.979 ± 0.009	0.993 ± 0.004	0.932 ± 0.016	0.967 ± 0.008	0.899 ± 0.048
	Precision	0.979 ± 0.011	0.989 ± 0.009	0.920 ± 0.028	0.965 ± 0.008	0.863 ± 0.075
	Recall	0.986 ± 0.007	0.996 ± 0.002	0.945 ± 0.019	0.969 ± 0.011	0.942 ± 0.038

uneven illumination (Smith et al., 2015) of partial images in actual experiments. Noise makes the boundaries of brain regions more difficult to identify. In this section, we specially selected datasets with noises to verify the robustness of our method. We added some of these noisy samples into train set; then, after training, we predicted testing datasets. For HIP-pg and CBX-gl, **Figure 7** shows the original images (A,D), the predicted binarized images (B,E) and the superpositions of images, the located Mask, and the predicted boundaries (C,F). The binarized images and the superposition images demonstrate that the localization and segmentation results on noisy images had the same good effect as on data without noise. Furthermore, **Figures 7G–J** show the details, illustrating that the segmented lines were well matched with the real boundaries, even in areas where the intensity difference was not obvious.

Applicability of DeepBrainSeg for Other Types of Datasets

We validated the effectiveness, accuracy, and robustness of DeepBrainSeg using the BPS datasets. To further illustrate the applicability, we present the segmentation results for other types of data from: MOST, MRI, and STP systems. For datasets from these three imaging systems, we selected HIP-pg and CBX-gl, caudoputamen (CP) and hippocampus (HIP), corpus callosum (CC) and HIP, respectively, for segmentation. In **Figure 8**, the first three rows show both the original images and the superposition images with the located Mask and the segmented lines from each of the three systems. DeepBrainSeg was able to effectively segment the brain regions from multiple types of datasets. The fourth row shows enlarged images of the areas in white boxes (**Figures 8A–F**). The detail images reveal that the segmented lines closely matched the real boundaries, indicating the wide applicability of our method.

Comparison With Other Methods

In this section, we compared DeepBrainSeg with other widely used methods including U-net, V-net, FC-DensNet, and Segnet. All methods were applied to BPS datasets with the same 60 training images and test images, and CBX-gl was selected as a

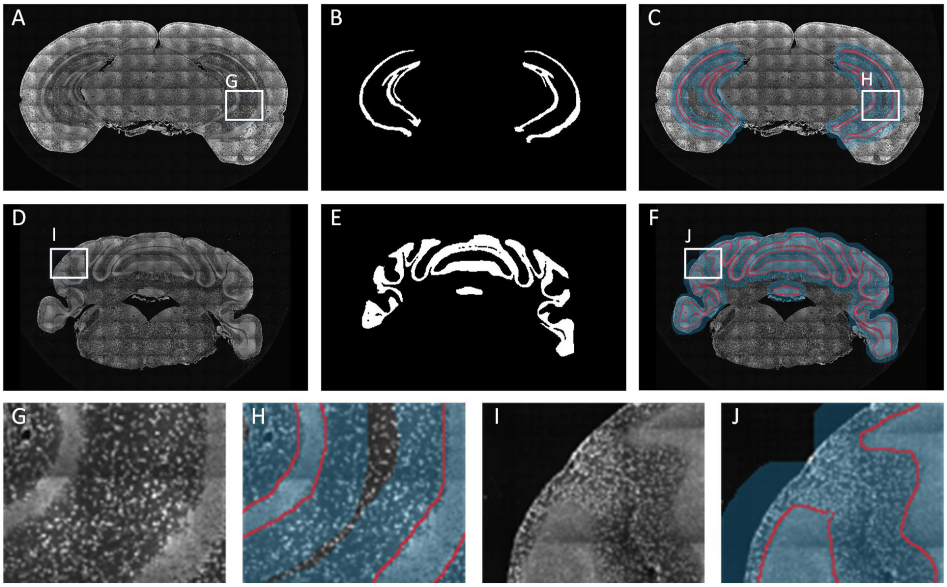


FIGURE 7 | Segmentation effects for datasets with noises. **(A–C)** The coronal image, the predicted binarized image, the superposition of image, the localization masks, and the segmented lines for HIP-pg. **(D–F)** The same as A-C for CBX-gl. **(G–J)** Enlarged views of the areas in white boxes in **A**, **C**, **D**, and **F**.

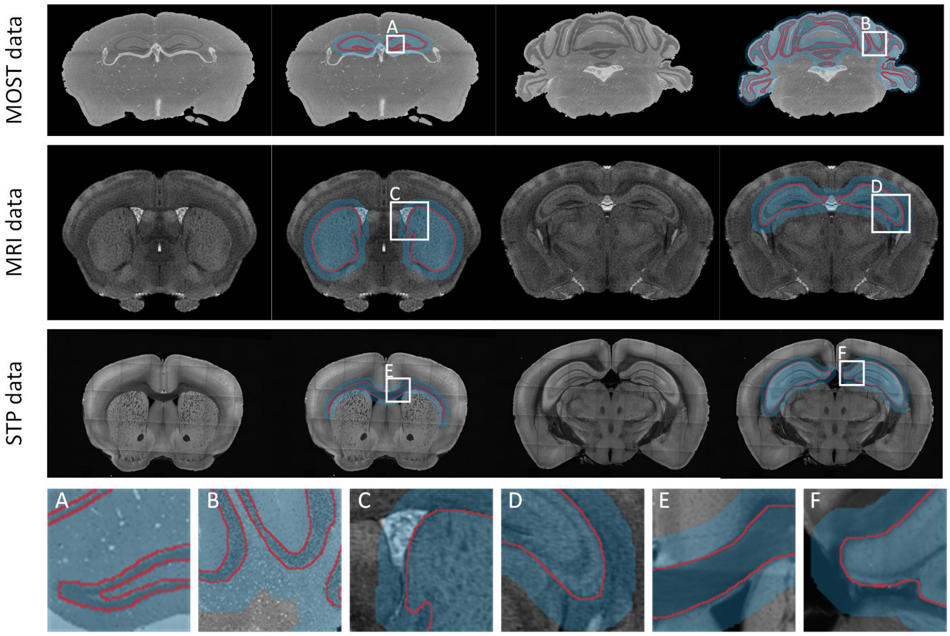


FIGURE 8 | Applicability of DeepBrainSeg for other types of datasets. The first row shows the segmentation effects of HIP-pg and CBX-gl for MOST data. From left to right: the coronal image, the superposition of image, the localization masks, and the segmented lines for HIP-pg and CBX-gl. The second and third rows show CP and HIP for the MRI data, CC, and HIP for STP data, respectively. **(A–F)** Enlarged views of the areas in white boxes in the first three rows.

representative structure with which to compare segmentation effects. The input images for DeepBrainSeg, U-net, V-net, and Segnet were full resolution of around 600 × 1000 pixels^2, while for FC-DensNet, they are limited to 400 × 600 pixels^2 due to the memory capacity of a GPU. **Figure 9** shows the results of these methods from top to bottom. The green lines indicate the ground truth by manual segmentation, and the red lines are the automatically segmented lines. The second, fourth, and fifth columns are enlarged images of the white boxes in preceding columns. These results show that other methods achieved general segmentation effects: some over-segmentation and erroneous segmentation were present

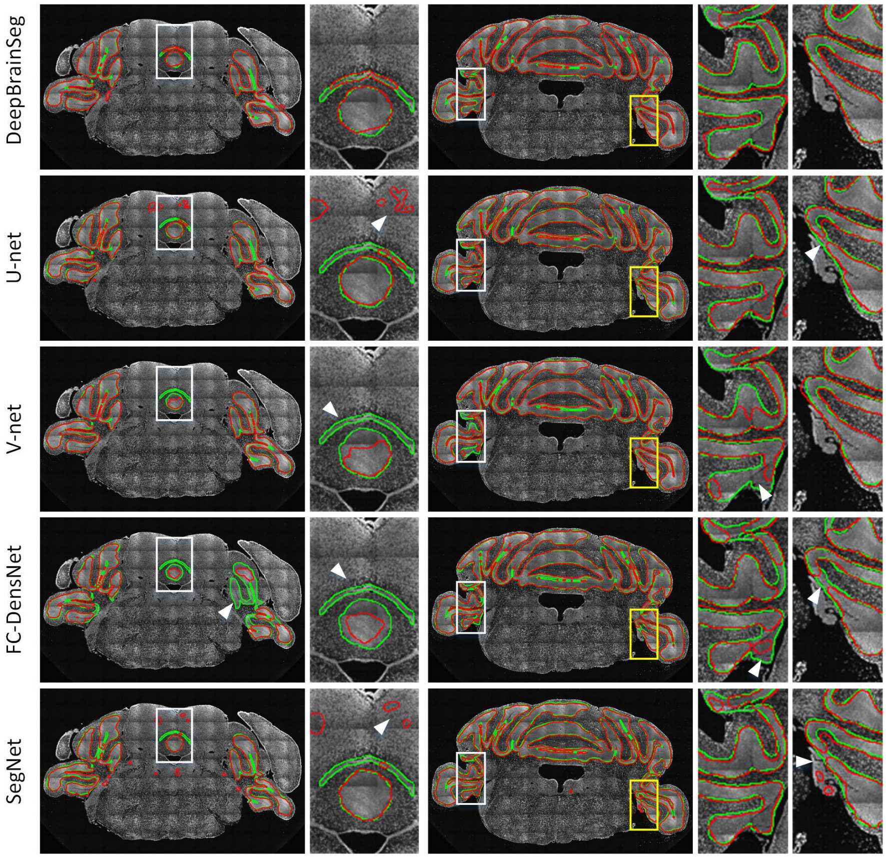


FIGURE 9 | Comparison among DeepBrainSeg, U-net, V-net, FC-DenseNet, and SegNet. From top to bottom, the five rows show the segmentation effects of these methods, respectively. The first and third images in each row are superpositions of coronal images and the segmented lines: the green lines are the ground truth and the red lines are the automatically segmented lines. The second, fourth, and fifth images show enlarged views of the areas in front white boxes. White arrows show the inaccurate segmentations.

TABLE 3 | Quantitative comparison among DeepBrainSeg and other methods (bold values are the highest).

	Dice	Precision	Recall
DeepBrainSeg	**0.960 ± 0.006**	**0.953 ± 0.009**	**0.968 ± 0.010**
U-net	0.929 ± 0.010	0.927 ± 0.019	0.931 ± 0.011
V-net	0.941 ± 0.007	0.929 ± 0.012	0.954 ± 0.016
FC-DensNet	0.919 ± 0.048	0.923 ± 0.014	0.922 ± 0.091
SegNet	0.911 ± 0.009	0.918 ± 0.018	0.904 ± 0.012

in the latter (marked by white arrows). In contrast, the segmented lines from DeepBrainSeg match more accurately with manual lines, and contain less erroneous segmentation. This indicates that DeepBrainSeg has a stronger segmentation ability for brain regions.

We also quantitatively evaluated the effects of the three methods in the test data. **Table 3** shows the mean and standard deviation values of Dice, Precision, and Recall. Our proposed method achieves the highest values for the three parameters: 0.960, 0.953, and 0.968, respectively. In addition, we conducted statistical tests for evaluated values by conducting Wilcoxson test between DeepBrainSeg and others. The *P*-values displayed in **Table 4**, Dice, Precision, and Recall values of DeepBrainSeg are significantly different from all others ($P < 0.05$).

TABLE 4 | *P*-values of Wilcoxson test among DeepBrainSeg and other methods.

	Dice	Precision	Recall
U-net	1.82e-04	7.69e-04	2.46e-04
V-net	3.30e-04	7.68e-04	3.12e-02
FC-DensNet	1.83e-04	1.00e-03	5.80e-03
SegNet	1.82e-04	2.46e-04	1.83e-04

Performance Evaluation

Benefiting from the optimization of the domain-condition constraint and prediction at intervals, our method significantly improved the computational efficiency. Ten consecutive coronal planes for five brain regions were selected to evaluate the number of pixels requiring computation before and after optimization, respectively. As shown in **Figure 10A**, when predicting each pixel in the entire

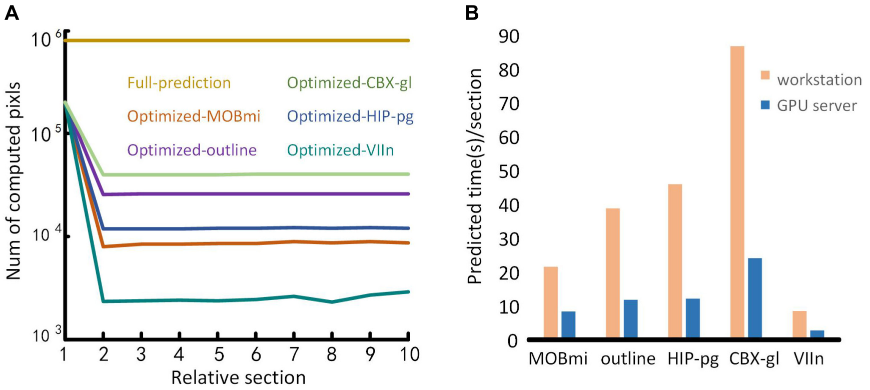


FIGURE 10 | Performance testing. **(A)** Comparison of the time required for full-image predictions and optimization predictions. The abscissa represents a sequence of ten consecutive coronal images. The ordinate is the number of pixels to be calculated. Different color lines represent the calculation required for different brain regions using full-image predictions and the optimization predictions. **(B)** Performance of the proposed method on different computing platforms. The abscissa represents the five brain regions, and the ordinate is the average prediction time for each image. The orange and blue bars represent the performances of the workstation and the GPU server platform, respectively.

image, the amount of calculation approaches 10^6 for the full prediction. In contrast, using the optimization method, the first image requires three times less calculation. For subsequent images, only the pixels in the mask needed to be predicted, the amount of calculation decreased by one to three orders of magnitude according to the size of different brain regions.

To evaluate the performance of our method on different computing platforms, we tested five brain regions on a graphical workstation with a M6000 GPU and on a GPU server with four V100 GPUs. The prediction time of each section for the five brain regions on the platforms is shown in **Figure 10B**. The maximum runtime of one section was 90 s on the workstation. Furthermore, the time for that section decreased approximately threefold when executed on the GPU server platform.

DISCUSSION

In this study, following the principle of manual segmentation with multi-level features, we proposed DeepBrainSeg to solve the issue of brain region segmentation for micro-optical images based on a CNN. We used a dual-pathway CNN to learn local and contextual information at different scales, and provided global localization through image registration and domain-condition constraints. Our method can accurately segment multiple brain-wide regions, even for datasets with noises, and is widely applicable to various types of datasets. Moreover, it is superior to U-net, V-net, FC-DensNet, and Segnet in terms of accuracy.

We demonstrated the segmentation effects of our method on four different types of data. Furthermore, DeepBrainSeg can also be applied to solve segmentation problems in other fields for more types of data, such as computed tomography and electron microscopy. For other data, the patch size and network structure require adjustment according to the ratio of its resolution to 10 μm. Meanwhile, the potential regions for segmentation are not limited to the examples shown in this paper: the method is also suitable for other regions with characteristic differences to their surroundings. For brain region localization, DeepBrainSeg provides a location area that is consistent with the shape of the real brain region, rather than a regular shape like box. This irregular location area reduces the Redundancy to improve the localization accuracy and segmentation efficiency.

Nevertheless, our method still has some deficiencies. The training and prediction are implemented separately that target the characteristics of these different brain regions but introduce some complexity. Thus, finding one model that can segment multiple brain regions will be the subject of our future work. In addition, for efficiency, we processed datasets at an isotropic resolution of 10 μm. It is likely that a higher resolution could achieved by improving the algorithm and efficiency.

Research for brain space information involves collaborative analysis of various brain regions and datasets. Although many methods have been applied for brain segmentation, they are generally effective for only one type of data or a single brain region. Our intention is to provide neuroscientists with a consistently accurate segmentation framework that can be applied to multiple types of data and brain regions without requiring complex feature extraction or being subject to strict data-quality requirements. Users would only need to input the data into the method to quickly acquire satisfactory results. We believe that our method provides a powerful tool by which neuroscientists can explore the brain.

DATA AVAILABILITY STATEMENT

The image data and codes supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The animal study was reviewed and approved by the Institutional Animal Ethics Committee of Huazhong University of Science and Technology.

AUTHOR CONTRIBUTIONS

QL and HG conceived the project. CT and AL designed the method. CT, AL, and YG wrote the article. CT, ZF, HN, ZZ, and ZW processed the data sets. XL prepared the brain specimens. JY processed the brain-wide imaging.

FUNDING

This work was supported by the Science Fund for Creative Research Group of China (Grant No. 61721092), the National Natural Science Foundation of China (Grant Nos. 81871082 and 91749209), and the director fund of the WNLO.

ACKNOWLEDGMENTS

We appreciate X. Chai and Y. Di for data processing, T. Hu for constructive suggestions, and Britton Chance Center for Biomedical Photonics for the support in data acquisition.

REFERENCES

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Med. Image Anal.* 12, 26–41. doi: 10.1016/j.media.2007.06.004

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *Paper Presented at the IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39 (Piscataway, NJ: IEEE), 2481–2495. doi: 10.1109/tpami.2016.2644615

Balafar, M. A., Ramli, A. R., Saripan, M. I., and Mashohor, S. (2010). Review of brain MRI image segmentation methods. *Artif. Intell. Rev.* 33, 261–274. doi: 10.1007/s10462-010-9155-0

Chen, H., Dou, Q., Yu, L., Qin, J., and Heng, P.-A. (2018). VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. *Neuroimage* 170, 446–455. doi: 10.1016/j.neuroimage.2017.04.041

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *Paper Presented at the IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 40 (Piscataway, NJ: IEEE), 834–848. doi: 10.1109/tpami.2017.2699184

Chen, Y., McElvain, L. E., Tolpygo, A. S., Ferrante, D., Friedman, B., Mitra, P. P., et al. (2019). An active texture-based digital atlas enables automated mapping of structures and markers across brains. *Nat. Methods* 16, 341–350. doi: 10.1038/s41592-019-0328-8

Ciresan, D., Giusti, A., Gambardella, L. M., and Schmidhuber, J. (2012). "Deep neural networks segment neuronal membranes in electron microscopy images", in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Vol. 2 (Red Hook, NY: Curran Associates Inc.), 2843–2851.

Clarke, L., Velthuizen, R., Camacho, M., Heine, J., Vaidyanathan, M., Hall, L., et al. (1995). MRI segmentation: methods and applications. *Magn. Reson. Imaging* 13, 343–368.

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* (Boston, MA: IEEE), 20–28.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology* 26, 297–302. doi: 10.2307/1932409

Dong, H. W. (2008). *The Allen Reference Atlas: A Digital Color Brain Atlas of the C57Bl/6J Male Mouse*. Hoboken, NJ: John Wiley & Sons Inc.

Economo, M. N., Clack, N. G., Lavis, L. D., Gerfen, C. R., Svoboda, K., Myers, E. W., et al. (2016). A platform for brain-wide imaging and reconstruction of individual neurons. *Elife* 5:e10566.

Feng, Y., Zhao, H., Li, X., Zhang, X., and Li, H. (2017). A multi-scale 3D Otsu thresholding algorithm for medical image segmentation. *Digit. Signal Process.* 60, 186–199. doi: 10.1016/j.dsp.2016.08.003

Gahr, M. (1997). How should brain nuclei be delineated? Consequences for developmental mechanisms and for correlations of area size, neuron numbers and functions of brain nuclei. *Trends Neurosci.* 20, 58–62. doi: 10.1016/s0166-2236(96)10076-x

Gong, H., Xu, D., Yuan, J., Li, X., Guo, C., Peng, J., et al. (2016). High-throughput dual-colour precision imaging for brain-wide connectome with cytoarchitectonic landmarks at the cellular level. *Nat. Commun.* 7:12142.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice, VE: IEEE), 2961–2969.

Iqbal, A., Khan, R., and Karayannis, T. (2019). Developing a brain atlas through deep learning. *Nat. Mac. Intell.* 1:277. doi: 10.1038/s42256-019-0058-8

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., and Bengio, Y. (2017). "The one hundred layers tiramisu: fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Honolulu, HI: IEEE), 11–19.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia* (New York, NY: ACM), 675–678.

Johnson, G. A., Badea, A., Brandenburg, J., Cofer, G., Fubara, B., Liu, S., et al. (2010). Waxholm space: an image-based reference for coordinating mouse brain research. *NeuroImage* 53, 365–372. doi: 10.1016/j.neuroimage.2010.06.067

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems* (New York, NY: ACM), 1097–1105.

Kuan, L., Li, Y., Lau, C., Feng, D., Bernard, A., Sunkin, S. M., et al. (2015). Neuroinformatics of the allen mouse brain connectivity atlas. *Methods* 73, 4–17. doi: 10.1016/j.ymeth.2014.12.013

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444.

Lai, M. (2015). Deep learning for medical image segmentation. *arXiv* [Preprint]. arXiv:1505.02000.

Li, A., Gong, H., Zhang, B., Wang, Q., Yan, C., Wu, J., et al. (2010). Micro-optical sectioning tomography to obtain a high-resolution atlas of the mouse brain. *Science* 330, 1404–1408. doi: 10.1126/science.1191776

Li, X., Yu, B., Sun, Q., Zhang, Y., Ren, M., Zhang, X., et al. (2018). Generation of a whole-brain atlas for the cholinergic system and mesoscopic projectome analysis of basal forebrain cholinergic neurons. *Proc. Natl. Acad. Sci. U.S.A.* 115, 415–420. doi: 10.1073/pnas.1703601115

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.

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: IEEE), 3431–3440.

Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2016). "Understanding the effective receptive field in deep convolutional neural networks," in *Proceedings of the Advances in Neural Information Processing Systems* (Red Hook, NY: NIPS), 4898–4906.

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

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J., Išgum, I., et al. (2016). Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans. Med. Imaging* 35, 1252–1261. doi: 10.1109/tmi.2016.2548501

Nanthagopal, A. P., and Sukanesh, R. (2013). Wavelet statistical texture features-based segmentation and classification of brain computed tomography images. *IET Image Process.* 7, 25–32. doi: 10.1049/iet-ipr.2012.0073

Peng, J., Long, B., Yuan, J., Peng, X., Ni, H., Li, X., et al. (2017). A quantitative analysis of the distribution of CRH neurons in whole mouse brain. *Front. Neuroanat.* 11:63. doi: 10.3389/fnana.2017.00063

Pereira, S., Pinto, A., Alves, V., and Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging* 35, 1240–1251. doi: 10.1109/tmi.2016.2538465

Ragan, T., Kadiri, L. R., Venkataraju, K. U., Bahlmann, K., Sutin, J., Taranda, J., et al. (2012). Serial two-photon tomography for automated ex vivo mouse brain imaging. *Nat. Methods* 9, 255–258. doi: 10.1038/nmeth.1854

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

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing And Computer-Assisted Intervention* (Berlin: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* 19, 221–248.

Smith, K., Li, Y., Piccinini, F., Csucs, G., Balazs, C., Bevilacqua, A., et al. (2015). CIDRE: an illumination-correction method for optical microscopy. *Nat. Methods* 12, 404–406. doi: 10.1038/nmeth.3323

Xiong, B., Li, A., Lou, Y., Chen, S., Long, B., Peng, J., et al. (2017). Precise cerebral vascular atlas in stereotaxic coordinates of whole mouse brain. *Front. Neuroanat.* 11:128. doi: 10.3389/fnana.2017.00128

Yu, L., Yang, X., Chen, H., Qin, J., and Heng, P.-A. (2017). "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (Palo Alto, CA: AAAI Press), 66–72.

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.

BRIEF RESEARCH REPORT
published: 26 March 2020
doi: 10.3389/fnins.2020.00207

Using Deep Convolutional Neural Networks for Neonatal Brain Image Segmentation

Yang Ding[1†], *Rolando Acosta*[1†], *Vicente Enguix*[1], *Sabrina Suffren*[1], *Janosch Ortmann*[2], *David Luck*[1], *Jose Dolz*[3] and *Gregory A. Lodygensky*[1,4,5*]

[1] The Canadian Neonatal Brain Platform (CNBP), Montreal, QC, Canada, [2] Department of Management and Technology, Université du Québec à Montréal, Montreal, QC, Canada, [3] Laboratory for Imagery, Vision and Artificial Intelligence (LIVIA), École de Technologie Supérieure, Montreal, QC, Canada, [4] Department of Pediatrics, Sainte-Justine University Hospital Research Center, Montreal, QC, Canada, [5] Department of Pharmacology and Physiology, University of Montreal, Montreal, QC, Canada

Edited by:
Amir Shmuel,
McGill University, Canada

Reviewed by:
Kuo Men,
Chinese Academy of Medical Sciences and Peking Union Medical College, China
Refaat E. Gabr,
University of Texas Health Science Center at Houston, United States

****Correspondence:***
Gregory A. Lodygensky
ga.lodygensky@umontreal.ca

† These authors share first authorship

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *22 August 2019*
Accepted: *25 February 2020*
Published: *26 March 2020*

Citation:
Ding Y, Acosta R, Enguix V, Suffren S, Ortmann J, Luck D, Dolz J and Lodygensky GA (2020) Using Deep Convolutional Neural Networks for Neonatal Brain Image Segmentation. Front. Neurosci. 14:207. doi: 10.3389/fnins.2020.00207

Introduction: Deep learning neural networks are especially potent at dealing with structured data, such as images and volumes. Both modified LiviaNET and HyperDense-Net performed well at a prior competition segmenting 6-month-old infant magnetic resonance images, but neonatal cerebral tissue type identification is challenging given its uniquely inverted tissue contrasts. The current study aims to evaluate the two architectures to segment neonatal brain tissue types at term equivalent age.

Methods: Both networks were retrained over 24 pairs of neonatal T1 and T2 data from the Developing Human Connectome Project public data set and validated on another eight pairs against ground truth. We then reported the best-performing model from training and its performance by computing the Dice similarity coefficient (DSC) for each tissue type against eight test subjects.

Results: During the testing phase, among the segmentation approaches tested, the dual-modality HyperDense-Net achieved the best statistically significantly test mean DSC values, obtaining 0.94/0.95/0.92 for the tissue types and took 80 h to train and 10 min to segment, including preprocessing. The single-modality LiviaNET was better at processing T2-weighted images than processing T1-weighted images across all tissue types, achieving mean DSC values of 0.90/0.90/0.88 for gray matter, white matter, and cerebrospinal fluid, respectively, while requiring 30 h to train and 8 min to segment each brain, including preprocessing.

Discussion: Our evaluation demonstrates that both neural networks can segment neonatal brains, achieving previously reported performance. Both networks will be continuously retrained over an increasingly larger repertoire of neonatal brain data and be made available through the Canadian Neonatal Brain Platform to better serve the neonatal brain imaging research community.

Keywords: neonatal brain, brain segmentation, machine learning (artificial intelligence), convolutional neural network, T2-weighed MRI

INTRODUCTION

The magnetic resonance imaging (MRI) study of brain development since birth represents one of the crucial modern techniques to improve our understanding of developmental neuroscience and help identify the long-term links between brain injuries and respective developmental consequences. However, despite mature analytical methods to process adult human brain MRIs, analyses of brains during development and especially at the neonatal stage remain difficult as a result of isolated tools development and difficulty with data acquisition. The most important step before performing quantitative brain analyses is the tissue class segmentation of the brain. Neonatal brain medical imaging tissue type identification is especially challenging given its typically inverted T1/T2 tissue contrast compared to adults (Shroff et al., 2010). Moreover, the amount of high-quality public neonatal research neural MRI data sets is far rarer in comparison to adult neural MRI data, making training, development, and adoption of neonatal-specific brain segmentation approaches challenging. From our experience in attempting to implement majority of the open-source neonatal segmentation approaches at the Canadian Neonatal Brain Platform (CNBP)[1], many existing computer-vision-based solutions failed to generalize beyond the respective niche of privately held training data set. As part of our organizational mandates, CNBP aims to validate and provide a large variety of neonatal brain MRI processing approaches. In this article, we focused primarily on public-data-based open-source deep learning approaches in the context of neonatal brain tissue segmentation.

Recent years have witnessed an explosive growth in the number of deep learning methods – especially convolutional neural network (CNNs) – for many vision problems, such as classification (Krizhevsky et al., 2012), detection (Ren et al., 2015), and semantic segmentation (Long et al., 2015). These models are capable of learning highly complex patterns by stacking multiple layers of convolutions and non-linear operations, presenting impressive capabilities to learn abstract representations from raw structured data in a data-driven manner. Particularly, the medical field has greatly benefited from these deep models, which have become the *de facto* solution for many of these tasks in highly important fields, such as radiology, oncology, or neuroimaging (Litjens et al., 2017).

Despite the fast adoption of these models in medical imaging, there have been relatively few large-scale efforts to find the top performer in pediatric brain segmentation using standardized open data sets (Akkus et al., 2017). Two particularly large-scale relevant competitions are known to date: the 2012 Neonatal Brains Segmentation Challenge[2] and the 2017 iSeg 6-month Infant Brain Magnetic Resonance Imaging Segmentation Challenge[3], both hosted as part of the respective Medical Image Computing and Computer Assisted Intervention Society (MICCAI) conferences. Out of the two competitions, the 2017 competition was particularly relevant as most contestants used derivation of CNN architecture forgoing traditional computer vision techniques, and some top performers openly shared their network architecture designs and code bases.

Outside of the iSeg 2017 competition and its related publications (Wang et al., 2019), which focus on 6-month-old infants, there have been few other proposed deep-learning-based segmentation approaches in neonates, despite numerous applications in either older infants (Zhang et al., 2015) or adults (Chen et al., 2018). The only applied neural network approach to solve neonatal tissue segmentation to date is from Moeskops et al. (2016). They proposed an integrated segmentation pipeline that reportedly can handle data from neonates all the way to 70-year-old adults using mini-patch-based 2D convolution approaches while only requiring a single anatomical reference MRI to achieve a respectable Dice score of at least 0.8 across five different data sets.

The objective of the current study is to evaluate both LiviaNET (Dolz et al., 2018b) and HyperDense-Net (Dolz et al., 2018a) architectures for neonatal brain imaging data. While both networks have demonstrated good performance on relevant tasks, such as in subcortical brain segmentation and in 6-month-old infant brain imaging data with diminished T1/T2 contrasts (Wang et al., 2012), their performance on neonatal-specific data remains untested. We hypothesize with a high-quality data set and ground truth, such as those from the publicly available Developing Human Connectome Project (DHCP) first-release neonatal data set (Hughes et al., 2017), we can achieve performance comparable to what prior modified LiviaNET and HyperDense-Net achieved in the adult and 6-month-old infant brain challenges. We aim to retrain both networks using the DHCP data set to validate the generalizability and the suitability of these network architectures in segmenting MRI brain tissue classes of neonatal brain images.

METHODS

Experimental Data: Participants

The participants were infants born at term from the publicly available DHCP by Hughes et al. (2017). DHCP is the first open-access data release of brain images of 40 healthy neonates born at term who had an MRI shortly after birth (37–44 weeks of gestational age). With these data, we had access to both raw data and tissue segmentation ground truth, generated using DrawEM and complemented further via manual correction, for training and validations. Additional MRI data-acquisition-related information is included in **Supplementary Method** as well as the original publication.

Experimental Data: Preprocessing

The training input was preprocessed based on the source image provided as part of the DHCP data made available (Hughes et al., 2017), namely, magnetic resonance bias-field correction with the N4 algorithm (Tustison et al., 2010) as implemented in Slicer 4.10.1 on our computational platform (see **Implementation: Computation Platform** section), launched with the command "Slicer – launch N4ITKBiasFieldCorrection."

[1] www.cnbp.ca

[2] https://neobrains12.isi.uu.nl/

[3] http://iseg2017.web.unc.edu/

Then the brain was extracted using the Brain Extraction Tool (BET2) with the default options (i.e., no additional customized command flags) from FMRIB Software Library (Smith, 2002; see **Figure 1A**). All T1-weighted images have been co-registered to the T2-weighted volumes using rigid alignment as implemented in SPM12 (Ashburner et al., 2014) in MATLAB (R2017b) (MathWorks Inc., Natick, MA, United States) running on our computational platform.

Experimental Data: Ground Truth Segmentation

As part of the DHCP data release, these neonatal brain MRIs were already segmented using the DHCP data pipeline built using the DrawEM module from the Medical Image Registration ToolKit (MIRTK) tool package (Makropoulos et al., 2014). DrawEM is an atlas-based segmentation technique that segments the volumes into 87 regions. Manually labeled atlases, annotated by an expert neuroanatomist (Gousias et al., 2012), were registered to the volume, and their labels were fused to the subject space to provide structure priors. Segmentation was then performed with an expectation–maximization scheme that combines the structure priors and an intensity model of the volume. The 87 regions were further merged to provide nine tissue segmentation labels provided with the DHCP release: (1) cerebrospinal fluid (CSF), (2) cortical gray matter (GM), (3) white matter (WM), (4) background, (5) ventricles, (6) cerebellum, (7) deep GM, (8) brain stem, and (9) hippocampus and amygdala. Since both LiviaNET and HyperDense-Net demonstrated their respective previous performance when dealing with four class labels (i.e., GM, WM, CSF, and others), we used the image calculator (ImCalc) function of SPM12 implemented in MATLAB R (2017b) (MathWorks Inc., Natick, MA, United States) to combine the existing nine DHCP class labels into the desired classes. More specifically, we combined together the cortical GM, cerebellum, deep GM, brainstem, and hippocampus and amygdala into the class "GM" and the CSF and ventricles into the class "CSF." The WM class was used as it is without change. What was originally left as the 10th class (i.e., unlabeled or outside) is considered as the fourth class (i.e., others). We included an illustration of an example subject in **Figure 1A** (top right).

Implementation: Network Architectures

In terms of network architectures, we evaluated two state-of-the-art networks that have shown outstanding performance for different brain segmentation tasks. The first network, referred to as LiviaNET (Dolz et al., 2018b), is a single-modality 3-D fully convolutional network which was proposed in the context of subcortical brain segmentation on MRI. At the time, standard segmentation convolutional neural networks performed slice-by-slice analyses of volumetric data. Nevertheless, an important limitation of this strategy is that the 3-D context orthogonal to the 2-D axial plane was completely discarded, resulting in segmentations without 3-D consistency. To address the computational and memory requirements of 3-D convolutions, LiviaNET adopted small kernels (27^3 voxels, **Figure 1A**, bounding box with green tab markers), resulting in deeper architectures with less complexity than their large-kernel counterparts. Furthermore, global and local contexts – important for both location and fine-grained details – were modeled by embedding intermediate-layer outputs in the final prediction. **Figure 1B** depicts the high-level architecture of LiviaNET.

The second network considered was HyperDense-Net (Dolz et al., 2018a), ranked among the top three methods in terms of performance in two different public data sets for adult (MRBRainS'13)[4] and isointense infant brain tissue segmentation (iSeg 2017)[5]. HyperDense-Net extends the previous network, LiviaNET, by leveraging dense connectivity in the context of multimodal image segmentation. Particularly, in this network, each image modality is processed in a different path, and dense connections occur between the pairs of layers within the same path, as well as across different paths. An example of this hyperdense connectivity is shown in **Figure 1C**.

Network parameters of both networks were optimized via a root mean square (RMS) optimizer (Hinton et al., 2014), using cross-entropy as a cost function to measure training error. This error was tracked throughout the training process and further elaborated in **Supplementary Method** along with additional network initialization parameters and hyperparameters.

Implementation: Experiment Design

There were 40 participants in total from DHCP data sets; they were split into three distinct groups: 60% of the subjects were for *training* (24 subjects), 20% were for *validation* to provide feedback on the neural network parameter tuning during training (eight subjects), and 20% were held out independently as the final *test* on the best-trained network to evaluate its generalization performance (eight subjects).

All subjects were randomly assigned to one of the three groups. The composition of the groups remains consistent throughout all experiments in both LiviaNET and HyperDense-Net network architectures.

Both networks were trained for a duration of 30 epochs composed of 20 subepoch each. At each subepoch, a total of 1,000 training subsamples (each composed of 27^3 voxel cubes, averaging about 41 samples per subject per setting) were randomly selected and given to the network, with a batch size of two.

At the end of the 30 epochs of *training*, the best-performing model as indicated by the *validation* data sets was evaluated on the holdout *test* data set in order to report the final test Dice similarity coefficient (DSC) values.

Implementation: Computation Platform

All training and testing were done using an Ubuntu 18.04 LTS running on a Xeon CPU E5-2600 Processor with 12 cores running at 2.0 GHz with 32 GB CPU DDR3 1,600 MHz RAM with a GeForce 1070 GPU with 8 GB of GDDR5 memory. Both HyperDense-Net and LiviaNET were implemented in Python

[4] https://mrbrains13.isi.uu.nl/

[5] http://iseg2017.web.unc.edu/

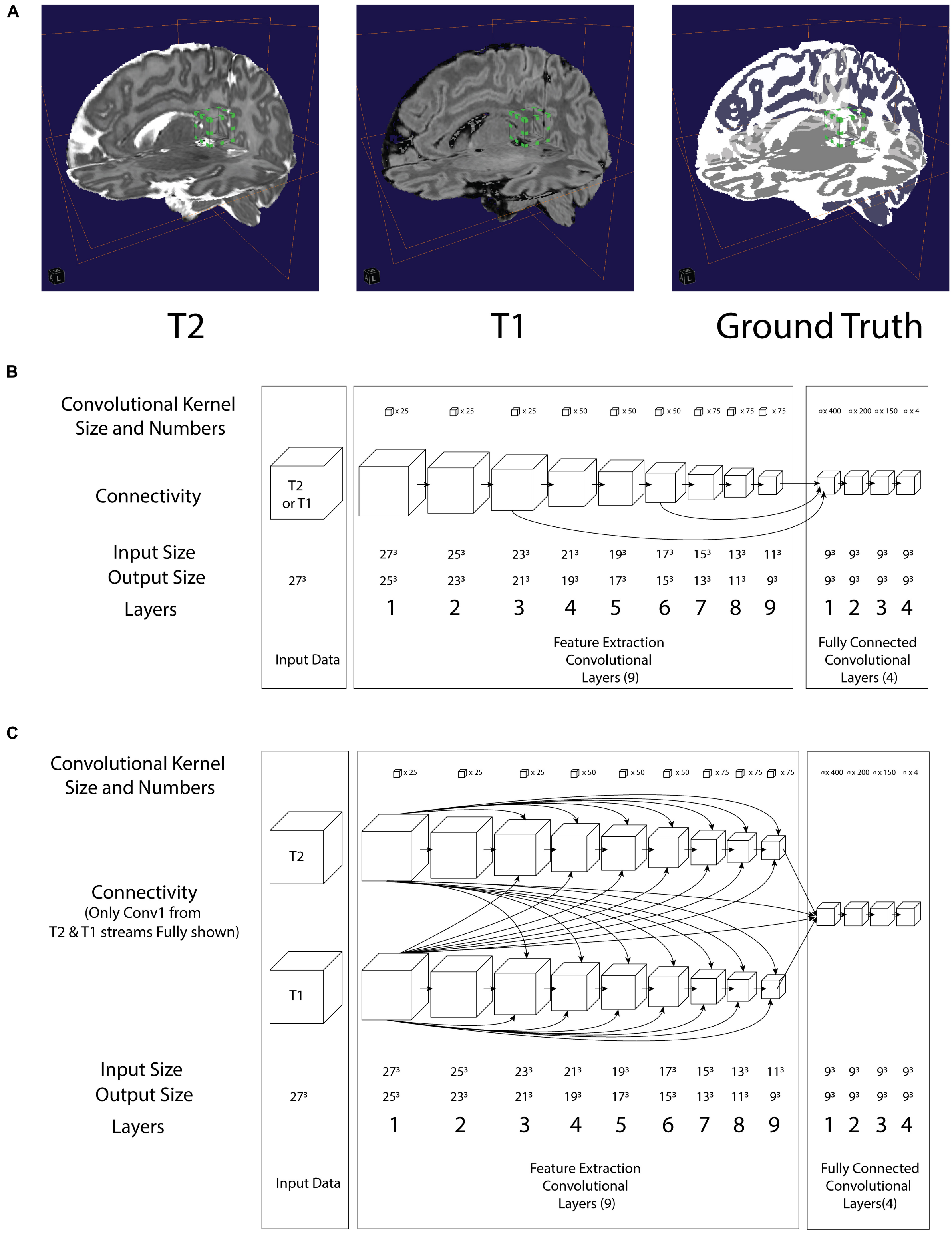


FIGURE 1 | (A) Illustration of a 3D convolution regional input (27 pixels^3) to both neural networks in relation to T2, T1 and Ground truth. **Bounding box with green tab:** input volume to the network **(B)** architecture of the LiviaNET illustrating major layer wise connections along with key parameters **(C)** architecture of HyperDense-Net neural network architecture including key parameters.

2.7 with Theano 1.0.0 library as per their source repositories at GitHub[6],[7].

Performance Evaluation

The DSC was used as the metric of final performance evaluation and computed separately in GM, WM, and CSF. In the context of tissue classification problem, it is an objective measure of both correctly classifying voxels of tissue where it belongs and correctly rejecting the voxels of incorrect tissue types.

The DSC is also known as the *Sørensen–Dice* coefficient or F1 score. DSC ranges between 0 and 1 with the perfect performance scored as 1. Its derivation and references are further elaborated in **Supplementary Method**.

Python 3.7 stats module was used to conduct pairwise T-tests to compare performance metrics from the same subjects during the prediction test against ground truth across various combinations of network architecture and data. Pairwise T-tests were also used for inter- and intra-architectural comparisons across epochs. Bonferroni correction was applied where appropriate to ensure the family-wise error rate is constrained to below 0.05. Jupyter notebook 1.0.0 and Plotly 4.0.0 library (Plotly, Montreal, Canada) were used to plot all figures in vector format before they were touched up in Adobe Illustrator CC 2017 (Adobe Systems Incorporation, San Jose, United States) for readability and DPI compliance formatting.

RESULTS

Training Performance

The final model of LiviaNET using T1 achieved a stable cross-entropy cost error of about 0.47 after approximately three epochs (**Figure 2**, row 1, left). When undergoing the same training settings but using only the T2 acquisitions, we achieved a cross-entropy cost error of 0.33 around a similar time point, which then remained consistent until the end of the training (**Figure 2**, row 1, middle). The final model weights of the HyperDense-Net achieved a relatively stable cross-entropy cost error of 0.24 after almost half way into the training process and experienced a much more gradual reduction of the standard deviation of cross-entropy cost error than LiviaNET (**Figure 2**, row 1, right). LiviaNET T2 and HyperDense-Net appear to have demonstrated reduced standard deviation of DSC during training compared to LiviaNET T1 across tissue types (**Figure 2**, rows 2–5). In addition, the superimposed trace (without standard deviation for clarity) of training cost error (**Supplementary Figure S2**) and of average DSC (**Supplementary Figure S3**) over training epochs was summarized in the same chart to facilitate comparisons of performance across architectures sharing both time and performance axes.

Test Performance

At the end of the training, the performance of the best model was tested against previously unseen eight holdout subjects' data as shown in grouped boxplots in **Figure 3**. The combination of LiviaNET and T1 data showed optimal performance at the 19th epoch and when tested resulted in prediction DSC values (mean ± standard deviation) of 0.86 ± 0.02, 0.86 ± 0.04, and 0.82 ± 0.04 for GM, WM, and CSF, respectively. Similarly, the optimal epoch for LiviaNET with T2 data was the 25th epoch and resulted in DSC values of 0.90 ± 0.02, 0.90 ± 0.01, and 0.88 ± 0.03, respectively. After accounting for multiple comparison problems via Bonferroni correction, the results demonstrate that LiviaNET using T2 data outperforms LiviaNET using T1 data significantly in most tissue types except white matter. For HyperDense-Net, the 29th epoch reported the optimal performance DSC at 0.94 ± 0.01, 0.95 ± 0.01, and 0.92 ± 0.03 for each tissue type compare to all LiviaNET results. Detailed statistical pairwise comparison results of the test performance are also included (**Supplementary Table S1**).

Time Benchmark

Using the aforementioned computational platform with NVIDIA GTX1070 GPU, LiviaNET took nearly 30 h to train for T1 input data and about 31 h for T2 while requiring 8 min on average (including preprocessing time) to segment a novel neonatal brain T1 or T2 scan. On the other hand, HyperDense-Net took about 86 h to train with both T2 and T1 data. In this case, segmentation of new neonatal data set was performed in nearly 10 min (including preprocessing time).

Visual Comparison

The segmentation outputs were visually inspected for congruency and obvious mistakes. We have uploaded the eight holdout test subjects, including the preprocessed T1 and T2 volume and ground truth labels to the accompanying GitLab repositories[8]. The segmentation results as both binary classification masks and tissue probability map for each subject are available for LiviaNET T1, LiviaNET T2, and HyperDense-Net T2 and T1 weighted. **Figure 4** shows a representative view of the segmentation output from one of the holdout test subjects. As illustrated, LiviaNET T1 (**Figure 4**, fourth column) struggled to identify WM properly especially near the deep GM regions. Across all three rows of different view perspectives, LiviaNET T1 misclassified multiple WM regions as GM, resulting the messiest view visually, congruent with its lower DSC result. On the other hand, both LiviaNET T2 and HyperDense-Net T2 and T1 segmentations resulted in better tissue separation and provided a closer match to the ground truth.

Comparison With Previously Reported Performance

In **Supplementary Table S2**, the average DSCs across tissue types of the best results obtained from the present experiments, along with the ones reported in the previous implementations of it, were listed for illustrative purposes. Since only mean accuracy was reported with no standard deviation or raw results available, no statistical comparisons were made.

[6] https://github.com/josedolz/HyperDenseNet

[7] https://github.com/josedolz/LiviaNET

[8] https://gitlab.com/dyt811/M017-Results

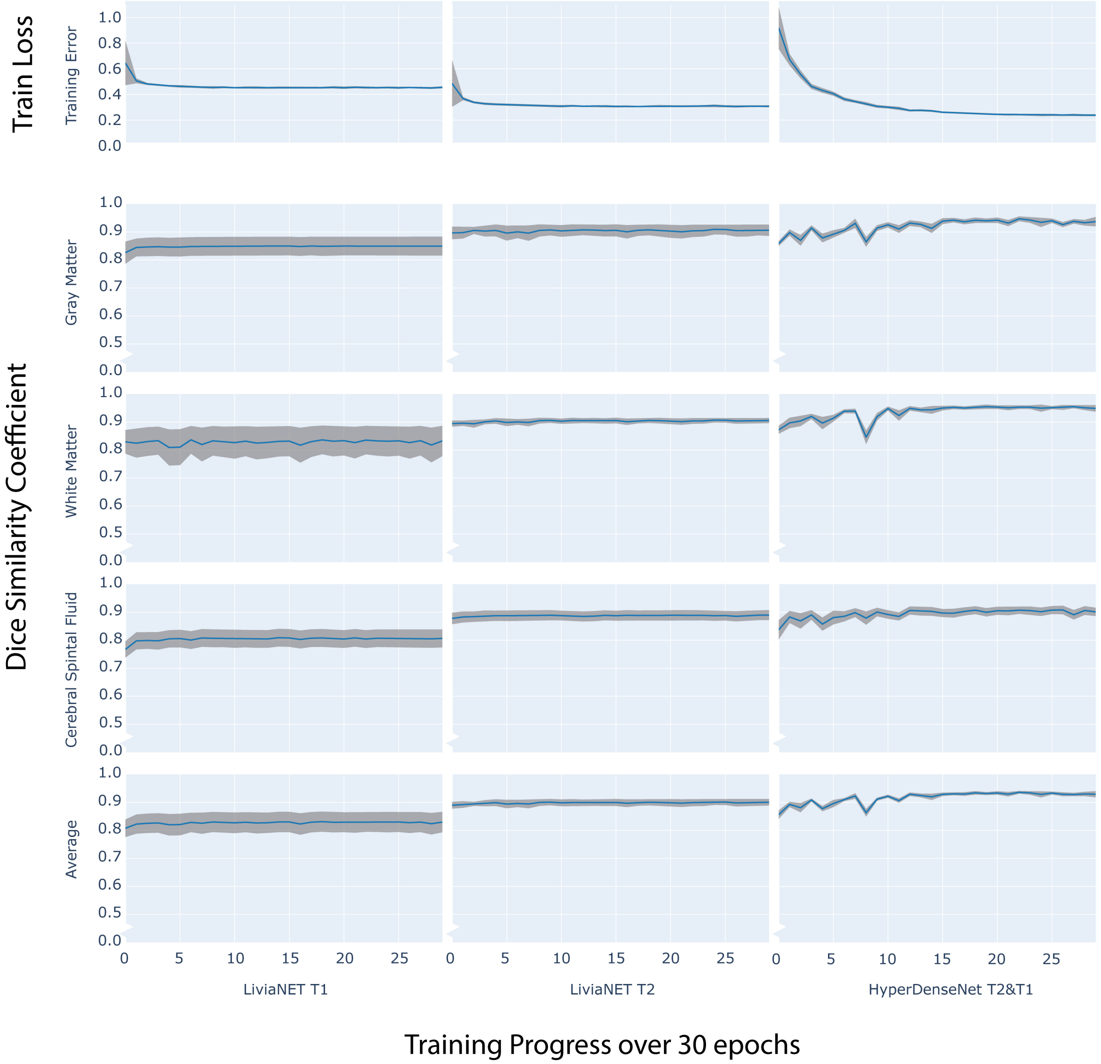


FIGURE 2 | Time series plot of over 30 training epochs measuring: training loss (top row) and Dice Similarity Coefficient (DSC) of Gray Matter (Row 2), White Matter (Row 3), Cerebrospinal Fluid (CSF Row4) and Average (Row 5) across LiviaNET using T1 (Column 1), T2 (Column 2), Hyperdense-Net using both T2 and T1 (Column 3). **Blue:** Mean measure across all eight test subjects. **Gray boundary:** standard deviation across all eight test subjects.

DISCUSSION

Summary

In this current work, both LiviaNET and HyperDense-Net architectures were evaluated using the publicly available DHCP neonatal data set. We demonstrated for the first time that the dual-modality HyperDense-Net performed significantly better in the context of neonatal brain segmentation specifically across all tissue types versus the single-modality LiviaNET. In addition, LiviaNET segments the neonatal brain better with T2-weighted images than with T1-weighted images.

Intramodel

LiviaNET has been primarily employed for single-modality inputs (i.e., T1-weighted images or T2-weighted images). Our current empirical results applying it for segmentation of neonatal T1- and T2-weighted data showed that LiviaNET with T2 contrasts performed statistically better for segmentation in neonates (**Figure 3** and **Supplementary Table S1**). This is likely due to improved tissue contrast in neonatal T2 versus T1 and is not surprising given that neonates typically exhibit such tissue characteristics prior to the reduced contrast phase from 6 to 8 months from myelination over early development (Wang et al., 2012). This can also be observed readily in T1 and T2 raw neonatal data (**Figure 4**), as well as the greater high signal intensity regions observed in a simple histogram of voxel intensity plot (see **Supplementary Figure S1**). Lastly, visual inspection of the LiviaNET output for both T1 and T2 shows that clearly there are some deep WM which was misclassified as GM. We suspect this may be sites of early myelination (Deoni et al., 2012), resulting in altered contrasts

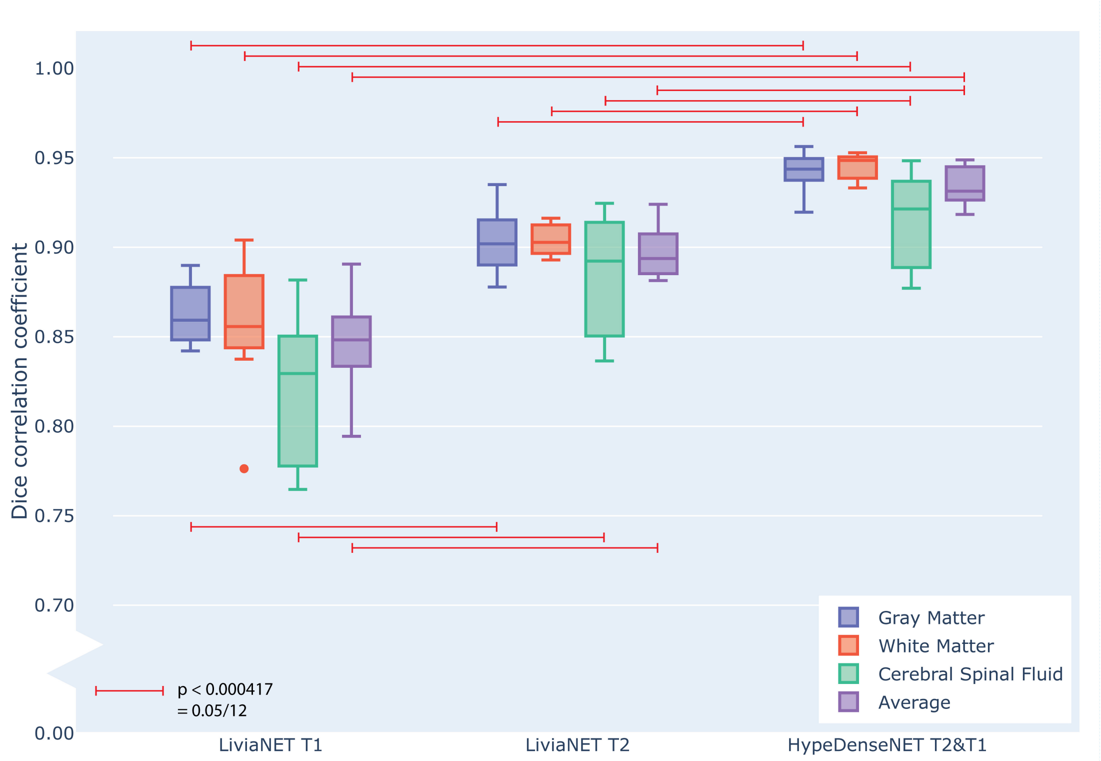


FIGURE 3 | Grouped box plot showing the Dice similarity coefficient (y axis) obtained during the testing phase across eight holdout subjects for each tissue type (color) and network types (x axis groups). Horizontal red lines denote family-wise error corrected statistically significant differences measured across the DSC in the same tissue using pairwise T statistical tests.

in comparison with surrounding tissues, which resulted in misclassification.

In terms of multimodal performance, HyperDense-Net was initially envisioned as a dual-/multi-modality version of LiviaNET, which derived its name from the extensive and dense connections between the T1 and T2 streams of successive convolutional layers. In this experiment, HyperDense-Net took longer to stabilize the training error across the eight validation (not test) subjects (**Figure 2**, row 1, right) and had also less stable DSC which fluctuated during training (**Figure 2**, rows 2–5) but eventually achieved relatively stable generalizable performance (**Figure 3**) midway through the training. This notably stronger variation during training and validation, yet still achieving excellent generalizable results, is likely attributed to the more interconnected complexity of the architecture, requiring more observations to fine-tune the model weights through back-propagation. The observed local fluctuations in validation accuracy is a common behavior when training deep neural networks (such as those seen around epoch 8, **Figure 2**, rows 2 and 3, right). During training, the network parameters are updated to optimize a training objective, based on training data, which does not guarantee that the parameter updates are optimal for the validation samples. This, together with a higher learning rate at the beginning of the training, increases the chances of having these local perturbations in the validation performance, particularly in an early stage of the training. Nevertheless, as long as the validation curve converges, these fluctuations are not considered as a problem. Indeed, there exist many works, including the original HyperDense-Net (see Figure 5 in the original HyperDense-Net paper Dolz et al., 2018a), which show that these fluctuations do not hamper the network performance.

Intermodels

All networks, regardless of design and data input type, achieved a reasonable test accuracy of higher than 80% in the independent holdout data set, and all required at least 1 day of GPU computation time to train effectively. As expected, both networks appear to benefit from the inclusion of T2-weighted images, potentially more so than from the inclusion of T1-weighted ones. This is likely due to the higher contrast found on T2-weighted images with respect to the T1-weighted ones for neonates (**Supplementary Figure S1**). This phenomenon is especially evident in LiviaNET-related experiments (**Figure 2**). Overall, the current explorative results across network architectures and data types suggest that HyperDense-Net utilizing both T2 and T1 data achieved the best statistically significant segmentation performance among all experiments (**Figure 3** and **Supplementary Figure S3**) despite requiring a substantial amount of training time (86 vs. 30 h).

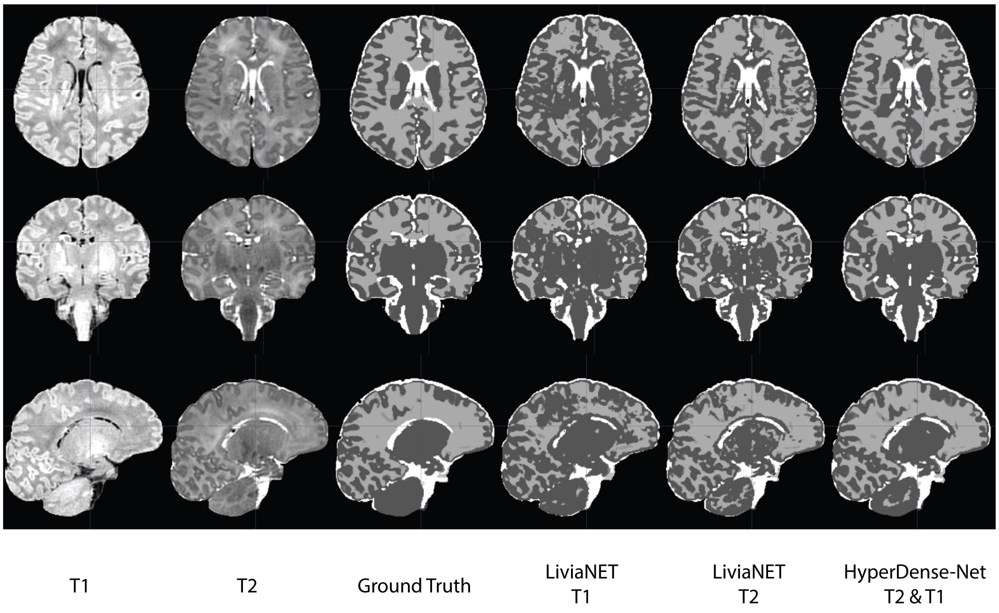


FIGURE 4 | Traverse ($z = 35$), coronal ($y = 5$) and sagittal ($x = 5$) slices of input data (T1-weighted, T2-weighted and ground truth tissue segmentation) registered to the final binary segmentation output of various networks trained (LiviaNET T1, Livia NET T2, HyperDense-Net T2&T1) on a single subject from the Developing Human Connectome Project (Subject CC00379XX17). **Crosshair** set at **MNI** coordinate of [5, 5, 35] and highlights the location of the respective slides from various views.

Compared to the modified LiviaNET version implemented for iSeg 2017 incorporating both T1 and T2 (Dolz et al., 2020), the current single-modality LiviaNET performance based on T2 data appears to be weaker consistently in the CSF classifications (**Supplementary Table S2**). Similarly, the current trained HyperDense-Net potentially performs on par or even slightly better in both GM and WM delineation while being worse in the CSF. Upon gross visual evaluation, we could not identify any major consistently common problems in the CSF relation regions, save for minor encroachment from the GM regions nearby. It might be necessary to conduct a spatial statistical parametric mapping type of analyses to truly evaluate the regions showing greater differences. However, given that we are observing this type of issues across network architectures and across data types, we suspect it might be rooted in the fundamental neonatal tissue MRI properties and should be further explored in more varied neonatal MRI acquisitions in the future.

Compared to the original HyperDense-Net training accuracy and mean DSC plot (see Figures 4, 5 from Dolz et al., 2018a), our current experiments with HyperDense-Net show similar if not slightly better and faster performance improvement from the original paper. We suspect this is also due to the improved tissue contrast at the neonatal stage versus 6-month infant data sets where onset of myelination starts to reduce the tissue contrast. Current neonatal data sets are all pre-myelination and hence may provide more information for the neural network, to better delineate tissue types, and result in faster learning and earlier observance of the performance-plateauing phenomenon. Another plausible explanation is related to the fact that for DHCP data input and ground truth, the inputs have all been preprocessed to remove non-brain-related tissues (via the Brain Extraction Tool) and to correct for non-homogeneity (N4), which could have substantially simplified the neural network's computation effort, as the bulk of the voxels within the 3-D acquisition volumes is likely non-brain tissue.

Performance Comparison

In terms of prediction speed, HyperDense-Net segmentation when applied to novel data was relatively fast. Although current hardware platform during the testing phase required about 8 min per participant for this study, previous reports suggest it can be even faster at 2–6 min with better-performing work-station-level graphics card such as NVIDIA Tesla P100 (Dolz et al., 2018a). Compared to other known neonatal segmentation methods such as DHCP data analysis pipeline, which takes around 7 h per participant (Makropoulos et al., 2014), or the approximately 30 min run time required by the morphological adaptive neonate tissue segmentation (MANTiS) toolbox (Beare et al., 2016), the HyperDense-Net prediction time requirement is well within reason. However, it is important to note that both of the other two traditional pipelines also conduct more granular regional identifications while both LiviaNET and HyperDense-Net are mostly tested with 3–10 classes of segmentation goals in the past, despite them being capable of conducting additional class segmentation should the ground truth be available. Since neither DHCP analyses pipeline nor MANTiS was ever officially submitted to be validated against the iSeg 2017 challenge data set, their unbiased accuracy can only be compared in neonatal data sets such as DHCP. Such comparisons, although interesting, are beyond the scope of this paper and will be the focus of our future work when we extend both neural networks to conduct more anatomical regional labeling.

Limitation and Future Work

The fast-evolving field of computer vision has witnessed the development of many deep segmentation architectures since the seminal works such as FCN (Long et al., 2015) for the segmentation of color scenes and UNet (Ronneberger et al., 2015) for medical images. The choice of the networks analyzed in the current study is based on the competitive performance obtained in very related tasks and the public availability of their implementations. The purpose of this paper, however, is not to achieve the best performance on the task at hand but to demonstrate their reproducibility and usability for neonatal brain segmentation. We expect that this study will have a positive impact on the neuroimaging community toward the ever-widening adoption of these deep learning models in neonatal brain segmentation. Thus, future work will include more extensive evaluation of these and other state-of-the-art segmentation neural networks, to assess the neonatal brain segmentation problem. We aim to highlight efficient networks which can produce accurate and reliable segmentations while comparing them against existing traditional computer vision approaches.

In the context of comparing with the earlier works in neonatal brain segmentation, another important limitation to be considered is the limited sample size of high-quality labeled data. In the neonatal imaging world, high-quality labels coupled with high-quality medical imaging data are exceptionally rare. One of the other similar public neonatal data sets authors were aware of only consists of 10 subjects (Alexander et al., 2017). We also reviewed the subjects used in older studies in the neonatal field and found, for instance, that most of the past highly cited neonatal segmentation techniques applying traditional computer vision had tested their performance on a similar if not fewer number of subjects (Prastawa et al., 2005; Weisenfeld and Warfield, 2009). This trend persists even in more recent work as summarized in Moeskops et al. (2016, Tables 3, 4), with most studies restricted to very few subjects with no more than 20.

Regardless of sample sizes and technical solution approaches, generalization to new data is very important in the field of image segmentation, especially given the wide array of MRI contrasts possible and inter-scanner and inter-sequence variations across institutions. Current results reported are trained, validated, and tested on publicly available DHCP neonatal data, which has identical acquisition condition, scanner model, and manufacturer. Furthermore, deep-learning-based models are well known for their poor generalization capabilities on unseen data. This is particularly important in future translation of research to practice, where (1) there exists a shift between images acquired under different conditions and (2) the model needs to be retrained as these images become available. The most feasible solution to address this issue is to adopt a continual learning strategy. This approach consists on incrementally retraining deep models while avoiding any virtual loss of memory on previous seen data sets, which may not be available during retraining. This line of work will be further explored in the near future by leveraging the infrastructure of our Canadian Neonatal Brain Platform, which is currently in the progress of acquiring neonatal brain imaging data with diverse acquisition conditions from across Canada for researchers. Our final goal is to leverage such infrastructure to continuously improve the performance of networks through exposure to the ever-increasing amount of neonatal data that become available while allowing individual neonatal researchers without such infrastructures to continuously benefit from our centralized effort at retraining the neural networks to peak performance.

CONCLUSION

The current study compared how two related convolutional neural network architectures addressed the automatic tissue segmentation task on neonatal brain MRIs. Among all pathways tested, HyperDense-Net showed the best performance in neonatal MRI tissue classifications. A streamlined and continuously retrained version of this will be deployed in the Canadian Neonatal Brain Platform, and we will continuously measure its performance against other competing segmentation approaches and newer network architectures.

DATA AVAILABILITY STATEMENT

The analyzed results for this study can be found in the public GitLab repository at https://gitlab.com/dyt811/M017-Results.

AUTHOR CONTRIBUTIONS

YD and GL conceived and designed the study. VE obtained the public database and organized it for analyses. SS prepared the ground truth for training with the help of YD. RA adapted and trained the neural networks. RA debugged the network data pipelines with the help of JO and JD. YD performed the statistical analyses, created figures and tables, and wrote the first draft of the manuscript. JD, SS, and RA wrote sections of the manuscript based on their respective areas of expertise. DL, GL, and JD provided critical feedback and organizational improvement to the manuscript. All authors contributed to the final manuscript revision and had read and approved the final submitted version.

FUNDING

This work was supported by the Brain Canada Foundation (BC PSG2014-3469) and the CHU Sainte-Justine Foundation (Canada) that funded the Canadian Neonatal Brain Platform (https://cnbp.ca).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00207/full#supplementary-material

REFERENCES

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D. L., and Erickson, B. J. (2017). Deep learning for brain MRI segmentation: state of the art and future directions. *J. Digit. Imaging* 30, 449–459. doi: 10.1007/s10278-017-9983-4

Alexander, B., Murray, A. L., Loh, W. Y., Matthews, L. G., Adamson, C., Beare, R., et al. (2017). A new neonatal cortical and subcortical brain atlas: the melbourne children's regional infant brain (M-CRIB) atlas. *Neuroimage* 147, 841–851. doi: 10.1016/j.neuroimage.2016.09.068

Ashburner, J., Barnes, G., Chen, C., Daunizeau, J., Flandin, G., Friston, K., et al. (2014). *SPM12 Manual*. London: Wellcome Trust Centre for Neuroimaging.

Beare, R. J., Chen, J., Kelly, C. E., Alexopoulos, D., Smyser, C. D., Rogers, C. E., et al. (2016). Neonatal brain tissue classification with morphological adaptation and unified segmentation. *Front. Neuroinform.* 10:12. doi: 10.3389/fninf.2016.00012

Chen, H., Dou, Q., Yu, L., Qin, J., and Heng, P.-A. (2018). VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. *Neuroimage* 170, 446–455. doi: 10.1016/j.neuroimage.2017.04.041

Deoni, S. C., Dean, D. C. III, O'muircheartaigh, J., Dirks, H., and Jerskey, B. A. (2012). Investigating white matter development in infancy and early childhood using myelin water faction and relaxation time mapping. *Neuroimage* 63, 1038–1053. doi: 10.1016/j.neuroimage.2012.07.037

Dolz, J., Ayed, I. B., Yuan, J., Gopinath, K., Lombaert, H., and Desrosiers, C. (2018a). HyperDense-Net: a hyper-densely connected CNN for multi-modal image semantic segmentation. *IEEE Trans. Med. Imaging* 38, 1116–1126. doi: 10.1109/tmi.2018.2878669

Dolz, J., Desrosiers, C., and Ayed, I. B. (2018b). 3D fully convolutional networks for subcortical segmentation in MRI: a large-scale study. *Neuroimage* 170, 456–470. doi: 10.1016/j.neuroimage.2017.04.039

Dolz, J., Desrosiers, C., Wang, L., Yuan, J., Shen, D., and Ben Ayed, I. (2020). Deep CNN ensembles and suggestive annotations for infant brain MRI segmentation. *Comput. Med. Imaging Graph.* 79:101660. doi: 10.1016/j.compmedimag.2019.101660

Gousias, I. S., Edwards, A. D., Rutherford, M. A., Counsell, S. J., Hajnal, J. V., Rueckert, D., et al. (2012). Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. *Neuroimage* 62, 1499–1509. doi: 10.1016/j.neuroimage.2012.05.083

Hinton, G. S., Nitish, S., and Kevin, S. (2014). *Neural Networks for Machine Learning Lecture 6*. Toronto, ON: University of Toronto.

Hughes, E., Cordero-Grande, L., Murgasova, M., Hutter, J., Price, A., Gomes, A. D. S., et al. (2017). The developing human connectome: announcing the first release of open access neonatal brain imaging. *Paper Presented at the 23rd Annual Meeting of the Organization for Human Brain Mapping*, Kanada.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Advances in neural information processing systems, NIPS 2012* (New York, NY: Communications of the ACM), 1097–1105.

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

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: IEEE), 3431–3440.

Makropoulos, A., Gousias, I. S., Ledig, C., Aljabar, P., Serag, A., Hajnal, J. V., et al. (2014). Automatic whole brain MRI segmentation of the developing neonatal brain. *IEEE Trans. Med. Imaging* 33, 1818–1831. doi: 10.1109/TMI.2014.2322280

Moeskops, P., Viergever, M. A., Mendrik, A. M., De Vries, L. S., Benders, M. J., and Išgum, I. (2016). Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans. Med. Imaging* 35, 1252–1261. doi: 10.1109/TMI.2016.2548501

Prastawa, M., Gilmore, J. H., Lin, W., and Gerig, G. (2005). Automatic segmentation of MR images of the developing newborn brain. *Med. Image Anal.* 9, 457–466. doi: 10.1016/j.media.2005.05.007

Ren, S., He, K., Girshick, R., and Sun, J. (2015). "Faster r-cnn: towards real-time object detection with region proposal networks," in *Proceedings of the Advances in Neural Information Processing Systems 28, NIPS 2015*, Montreal, QC.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Berlin: Springer, 234–241. doi: 10.1007/978-3-319-24574-4_28

Shroff, M. M., Soares-Fernandes, J. P., Whyte, H., and Raybaud, C. (2010). MR imaging for diagnostic evaluation of encephalopathy in the newborn. *Radiographics* 30, 763–780. doi: 10.1148/rg.303095126

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Wang, L., Nie, D., Li, G., Puybareau, É, Dolz, J., Zhang, Q., et al. (2019). Benchmark on automatic 6-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. *IEEE Trans. Med. Imaging* doi: 10.1109/TMI.2019.2901712 [Epub ahead of print].

Wang, L., Shi, F., Yap, P.-T., Gilmore, J. H., Lin, W., and Shen, D. (2012). 4D multi-modality tissue segmentation of serial infant images. *PLoS One* 7:e44596. doi: 10.1371/journal.pone.0044596

Weisenfeld, N. I., and Warfield, S. K. (2009). Automatic segmentation of newborn brain MRI. *Neuroimage* 47, 564–572. doi: 10.1016/j.neuroimage.2009.04.068

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., et al. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage* 108, 214–224. doi: 10.1016/j.neuroimage.2014.12.061

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 Neuroscience

ORIGINAL RESEARCH
published: 30 April 2020
doi: 10.3389/fnins.2020.00417

Paired Trial Classification: A Novel Deep Learning Technique for MVPA

Jacob M. Williams[1]*, Ashok Samal[1], Prahalada K. Rao[2] and Matthew R. Johnson[3]

[1] Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE, United States, [2] Department of Mechanical Engineering, University of Nebraska-Lincoln, Lincoln, NE, United States, [3] Department of Psychology, University of Nebraska-Lincoln, Lincoln, NE, United States

Many recent developments in machine learning have come from the field of "deep learning," or the use of advanced neural network architectures and techniques. While these methods have produced state-of-the-art results and dominated research focus in many fields, such as image classification and natural language processing, they have not gained as much ground over standard multivariate pattern analysis (MVPA) techniques in the classification of electroencephalography (EEG) or other human neuroscience datasets. The high dimensionality and large amounts of noise present in EEG data, coupled with the relatively low number of examples (trials) that can be reasonably obtained from a sample of human subjects, lead to difficulty training deep learning models. Even when a model successfully converges in training, significant overfitting can occur despite the presence of regularization techniques. To help alleviate these problems, we present a new method of "paired trial classification" that involves classifying pairs of EEG recordings as coming from the same class or different classes. This allows us to drastically increase the number of training examples, in a manner akin to but distinct from traditional data augmentation approaches, through the combinatorics of pairing trials. Moreover, paired trial classification still allows us to determine the true class of a novel example (trial) via a "dictionary" approach: compare the novel example to a group of known examples from each class, and determine the final class via summing the same/different decision values within each class. Since individual trials are noisy, this approach can be further improved by comparing a novel individual example with a "dictionary" in which each entry is an average of several examples (trials). Even further improvements can be realized in situations where multiple samples from a single unknown class can be averaged, thus permitting averaged signals to be compared with averaged signals.

Keywords: EEG, MVPA, deep learning, machine learning, cognitive neuroscience

Edited by:
Dong Song,
University of Southern California,
United States

Reviewed by:
Zhenyu Liu,
Institute of Automation (CAS), China
Biao Sun,
Tianjin University, China

****Correspondence:***
Jacob M. Williams
jakew42@gmail.com

Specialty section:
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: *16 August 2019*
Accepted: *06 April 2020*
Published: *30 April 2020*

Citation:
Williams JM, Samal A, Rao PK and Johnson MR (2020) Paired Trial Classification: A Novel Deep Learning Technique for MVPA.
Front. Neurosci. 14:417.
doi: 10.3389/fnins.2020.00417

1. INTRODUCTION

Deep learning has produced state-of-the-art results in many areas of machine learning, but adoption of deep learning for the classification of electroencephalography (EEG) signals, and other types of human neuroscience datasets, has lagged compared to its popularity in other fields. Although an increasing number of studies are using deep learning to process neuroimaging datasets, the improvements in performance have typically not been as drastic as in other fields (Lotte et al., 2018), and most human neuroscience research has continued to use more traditional multivariate pattern analysis (MVPA) approaches: Manual feature extraction followed by a simple, typically linear, classifier, such as support vector machines (SVMs; Cortes and Vapnik, 1995) or logistic regression and its derivatives, e.g., sparse multinomial logistic regression (SMLR; Krishnapuram et al., 2005).

Nevertheless, deep learning techniques are being explored in EEG classification. Bashivan et al. (2015) used a recurrent convolutional model to classify EEG data that was projected onto a two-dimensional plane and then subjected to Fourier analysis. The final model achieved an error rate of 8.89%, as compared to a 12.59% error rate with a random forest. While this is a meaningful reduction in error rate, boosting was not employed in the training of the random forest, which likely would have significantly shrunk the difference in performance. Lawhern et al. (2018) explored the use of fully convolutional neural networks; they applied convolutions in data that were arranged in a (channels × timepoint) fashion to create a two-dimensional matrix. These models had very few features, on the order of 2,200. This work showed improvements over the Filter Bank Common Spatial Pattern algorithm in a majority of the datasets tested, including the P300 event-related potential (ERP) in an oddball task, error-related negativity in brain-computer interfaces, movement-related cortical potential in a finger movement task, and sensory motor rhythm in imagined movement. Schirrmeister et al. (2017) further demonstrated the applicability of convolutional neural networks in decoding raw EEG signals without hand-crafted features. They showed that the learned filters were able to extract information in the alpha, beta, and high gamma wavelengths, and found a small improvement over the Filter Bank Common Spatial Pattern algorithm in their test dataset (82.1% accuracy to 84.0%) accuracy.

There are many possible reasons for modern deep learning techniques to underperform in EEG classification, compared to the drastic benefits deep learning has had for other fields. For one thing, EEG data are very noisy. The electrical activity that makes it to the recording electrodes is spatially smoothed and otherwise distorted by passing through poor conductors, such as the skull and scalp. Signals propagating in opposite directions interfere with each other and reduce the signal that makes it to the sensor. Even more importantly, human subjects' cognition and brain activity naturally fluctuate from trial to trial; on some trials, they may not be focused on the task at all, and thus may produce brain signals that poorly reflect the trial type they are presented with. As such, if participants' attentiveness cannot be inferred from behavioral performance, some trials may not be classifiable at all, despite lacking any overt signal artifacts. While averaging can be used in some cases to reduce the impact of this unclassifiable data, it is not practical in all situations, such as when working with brain-computer interfaces where real-time, single-shot classification is the ultimate goal.

EEG data also have very high dimensionality. Signals in most cognitive neuroscience studies are generally recorded with a sampling rate of between about 250 and 1,000 Hz, with anywhere between about 16 and 256 channels of data. Overfitting is common on such high-dimensional signals. This problem is further exacerbated by the limited number of examples (trials)[1] that are usually available. It is impractical to collect EEG datasets on the scale of hundreds of thousands of examples, as seen in other deep learning applications, such as image classification, as this would require extraordinarily long recording sessions with human subjects and/or an unreasonably large number of them. Finally, this is all further compounded by the large individual differences between different human subjects (e.g., Valenzi et al., 2014). While a digital image of, say, a traffic light could be taken from many different angles, under many different lighting conditions, etc., traffic lights still have a number of visual properties that are presumed to be more-or-less invariant across different conditions and exemplars; if a so-called "traffic light" were shaped like a pyramid, gelatinous, and translucent, and contained lights of blue/magenta/orange, most image classifiers (including human beings) would fail to recognize it as such, but it could also rightly be argued that those changes make it no longer a true "traffic light" anyway. In comparison, it is much more difficult to make such distinctions in patterns of neuroscience data across human beings; while certain general phenomena appear to be near-universal across most humans, such as the N170 ERP to face stimuli (Bentin et al., 1996), there is still substantial variation across individuals and trials that can be sufficient to fool many classifiers. And, because it is usually impractical to determine whether these variations are due to differences in head shape, recording artifacts, fluctuating attention, functional brain organization/connectivity, cognitive strategy, etc., it is much more difficult to establish any kind of ground truth as to what an ideal response would look like. Suppose a human participant exhibits no N170 ERP but has intact face recognition ability, with no discernible artifacts in their data; how do we reconcile this? In the EEG data, we have the equivalent of a pyramidal, gelatinous "traffic light" but are confronted with the awkward task of trying to determine if we can possibly align it, somehow, with all the other pictures of rectangular solid ones.

Even if deep learning has not yet produced drastic improvements in classification performance relative to traditional MVPA techniques for most cognitive neuroscience applications, it is still worth exploring further; there are a tremendous number of possible configurations of deep neural network architectures, and thus far we have only scratched the surface of what might be possible with them. However, if we do want to increase performance in the analysis of neuroscience data with deep learning, it might be wise to begin thinking about ways of changing how we could reformulate the basic problem. This paper describes one such possible reformulation (out of probably

[1] The common parlance in psychology and cognitive neuroscience would be "trials," but the machine learning literature usually says "samples" or "examples." Given the confusion of using "samples" (since these can also refer to individual data points of EEG), we will use "trials" to refer to these trial-like chunks of neuroscience recording data, in which the intent is to classify each trial/example into one of several categories. One caveat is that our approach relies on combining two "trials" of EEG data to form each "example" for classification, so while "trials" and "examples" would be synonymous for most traditional classification schemes, they are semantically distinct in the context of our PTC approach. Thus, in the present paper, we use "trials" to denote a short chunk of EEG data to be classified, and "examples" to denote the units fed into a classifier, which are either individual "trials" in traditional approaches or pairs of "trials" in PTC.

many): Instead of classifying a single example at a time, one could instead attempt to classify *pairs* of examples as belonging to either the same class or different classes. We refer to this general approach as paired trial classification (PTC), described further in section 2.2.2. This method presents several potential benefits. First, it allows for a drastic increase in the number of training examples, as there are $O(n^2)$ possible pairs. This makes it easier to find a neural network model that reliably converges, which can be a significant issue in datasets with a comparatively large number of features but comparatively few examples. Also, given the otherwise low impact of standard data augmentation techniques in the field (Bashivan et al., 2015), PTC could also potentially improve the ability of the model to generalize to new data by reducing the likelihood of the model to memorize samples from the dataset. Second, it reduces the problem to two classes, potentially simplifying multi-class problems and thus presenting a second way of making it easier to achieve robust classification performance from limited training data. Third, it is flexible: The basic same/different judgment can be interpreted either categorically or continuously, as a kind of similarity metric; it can be combined with a "dictionary" approach (see below) to achieve traditional multiclass classification; and trained PTC models can in principle be used with any input data, not necessarily just the categories it was initially trained with, which could have interesting theoretical applications in the future.

As alluded to above, when trying to classify a novel example into one of several categories, PTC could still be used by employing a "dictionary" approach. That is to say, a new trial can be compared to known trials from all known classes and classified as the same class as the exemplar(s) to which it is/are most similar. Thus, for a single trained network model, this allows us to classify each example in the test set multiple times and average the results of those individual decisions into a single overall classification, which has the potential to reduce variability in classification performance for individual trials. Novel trials could also be compared against averaged signals from multiple trials drawn from the training set. This allows for more stability in the comparisons, further addressing the issue of noise in EEG signals. Similar approaches based on "dictionary" comparisons and/or averaging have been used with traditional MVPA going back to its neuroimaging roots (Haxby et al., 2001), but PTC allows those approaches to be combined with the power and flexibility of deep learning.

Ideas similar to PTC, also with an intent to increase the size of the dataset and the accuracy of the classifier, have been explored in other domains. A similar pairing technique has also been explored, but at the pixel-classification level, in hyperspectral imaging. Rather than classifying individual pixels, Li et al. (2016) classified a pixel in combination with each of its neighboring pixels and used a voting strategy to determine the class of the original pixel. Another similar approach by Inoue explored data augmentation through the unweighted averaging of two images in the training set. These images were not required to be drawn from the same class but were always given the label of the first chosen image, thus preventing perfect memorization of the data. The final fine tuning was performed on unaugmented data. They demonstrated substantial performance improvements on the ILSVRC 2012 and CIFAR-10 datasets using GoogLeNet. This approach differs from our proposed approach in that it performs averaging rather than concatenation, and does not attempt to predict the sameness of the two samples (Inoue, 2018). Using a technique termed "Matched Pair-Learning," Theiler (2013) paired two signals with statistical dependence but differing labels (e.g., different frames of chemical plume data) and classified the pair together, but their aim was not to classify whether those signals had the same or different category, *per se*. Comparisons of trials of neuroscience data to each other, or to averaged sets of trials, have also been relatively commonly applied in "pattern similarity" analyses within the traditional MVPA domain of cognitive neuroscience; in essence, our method is an enhanced version of those pattern similarity approaches, which would typically use Pearson correlation, Euclidean distance, or other distance/similarity metrics (for more, see section 2.2.2 below). However, to our knowledge, this is the first time such an approach has been tried within the domain of deep learning, or conversely, the first occasion in which deep learning has been applied in the domain of pattern-similarity MVPA to achieve a similarity metric customized to the dataset at hand, and thus one that is "smarter" than existing metrics based purely on mathematical formulas.

2. MATERIALS AND METHODS

2.1. Data

We used an EEG dataset consisting of the "initial presentation" period of a cognitive neuroscience study first published by Johnson et al. (2015); for full details, please see that paper. Briefly, in the pertinent portion of that study, participants were presented with a pair of visual stimuli for 1,500 ms: either two written words, two images of faces, or two images of scenes. Thirty-one channels of EEG data were recorded at 250 Hz. The signals were bandpass-filtered via hardware in a 0.01–100 Hz range, and recorded with 14 bits of precision. See **Figure 1** for an illustration of the stimuli and data.

A total of 37 subjects participated in the study and had high enough quality data to be used. We used the same initial pre-processing steps, trial rejection parameters, and participant exclusion criteria as described originally by Johnson et al. In the original publication, there were two experiments with $N = 21$ and $N = 16$, but the "initial presentation" period did not differ between experiments, and thus we have combined them into a single $N = 37$ dataset for present purposes. Each subject had ∼200 trials after artifact rejection (around 60–70 per image category), for a total of just under 7,000 trials.

The data were then subjected to additional pre-processing for the deep-learning-based PTC analyses. All data values (originally in raw microvolts) were divided by a fixed factor of 20 to bring their scale approximately into the −1 to 1 range in which neural networks perform the best. Additionally, time averaging was applied to reduce the dimensionality of the data by a factor of 10, i.e., data were downsampled into time bins of 40ms apiece, similar to the bins used for MVPA in the original publication (Johnson et al., 2015). Thus, the total data dimensionality per trial was $31\ \textit{channels} \times 37\ \textit{time bins} = 1,147\ \textit{features}$.

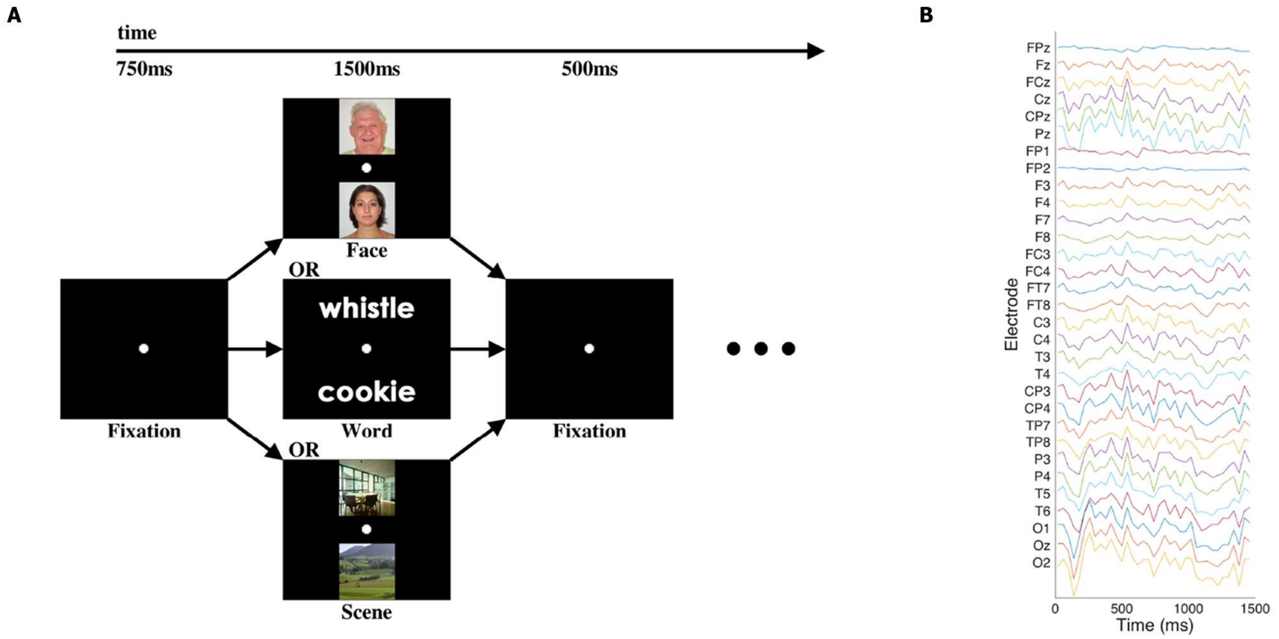


FIGURE 1 | Cognitive task and sample EEG data. **(A)** Participants viewed pairs of one of three categories of images at the beginning of each trial of the cognitive task, with blank-screen fixation intervals before and after. Other task components followed the presentation of the images, but those elements of the task are not presented or analyzed here. **(B)** Single representative trial of EEG data after pre-processing and downsampling. Electrode labels are according to the standard 10–20 and 10–10 systems for EEG electrode placement.

2.2. Classification Methodology

2.2.1. Baseline Models

In order to attain a baseline classification accuracy on the dataset, several widely used classifiers were examined. These include both a traditional classification baseline and a deep learning classification baseline. These models were trained on both single trials and averaged trials to allow for comparisons between the PTC methodology and other established techniques.

Traditional classification baselines were set using SVM and SMLR techniques, which are both frequently applied in traditional MVPA[2]. SVM analyses used a linear kernel, which is also common in neuroscience studies using MVPA, and which we have found to outperform radial basis function kernels in some of our previous analyses of EEG data. SVM hyperparameters were chosen with grid search over C in the set [0.0001, 0.001, 0.01, 0.1, 1, 10, 100]. Similarly, the lambda hyperparameter in SMLR was chosen through grid search from the set [0.001, 0.01, 0.1, 1, 10, 100, 1000]. These value ranges were chosen to span a range commonly seen in practice, in order to ensure that our comparisons were as fair as possible to the baseline conditions (i.e., that we did not hamstring the baselines by a poor choice of hyperparameter).

The deep learning model developed for paired-trial classification (see below) was also used for baseline analyses as a more conventional three-class neural network classifier (deep neural network [DNN] baseline model). This was done by slightly altering the network architecture, namely, by modifying the input layer to accept a single trial (rather than a pair), and modifying the output layer to have three output nodes instead of two, while leaving all hidden layers the same between the baseline DNN and PTC network architectures. It is certainly possible, given the effectively infinite number of combinations of architectures and hyperparameters, that better-performing DNN models could be found for the baseline analysis and/or the PTC analysis; however, for this initial demonstration of PTC we simply chose one relatively straightforward model that we thought would be fairly representative of the types of DNNs used to analyze cognitive neuroscience data.

2.2.2. Paired Trial Classification (PTC) Technique

The essence of the PTC approach is that instead of training a neural network to classify a single trial of data into one of several classes, the network is instead trained to determine whether two trials of input data are drawn from the same class or different classes. This binarization of the problem is somewhat different to the approach of, say, performing multiclass classification with SVMs by creating a number of binary SVMs and summing their outputs, because with the PTC approach, the same network can theoretically learn to classify similarities or differences between pairs of trials drawn from any class, for

[2] The method originally used for this dataset by Johnson et al. (2015) was SMLR, but in that paper, the authors were analyzing a different portion of the cognitive task and used a different cross-validation scheme that would not be readily comparable to our PTC approach or other baseline analyses in this paper, so those previous findings are not discussed here.

theoretically any number of classes. As such, PTC essentially gives us a new kind of similarity/distance metric with some useful characteristics: It can be interpreted either as a categorical same/different judgment or a continuous similarity/dissimilarity score, and it is "smarter" than simple formula-based metrics, such as Euclidean distance, cosine similarity, and Pearson correlation, having been trained to be sensitive to the features of a specific dataset that matter in differentiating the classes, while ignoring any nuisance features. To do this, each example fed into the classifier is comprised of two trials of EEG data (with dimensions Samples × Channels), stacked together to form a 2 × Samples × Channels input. Regardless of how many classes or conditions are present in the original dataset, the output layer always has two units, one representing a decision that the two trials in the input example are drawn from the same category, and the other unit representing a decision that the two trials are drawn from different categories. This also means that, in theory, a trained PTC network could be applied or adapted (via transfer learning) to previously unseen categories, although in this initial demonstration we do not yet test that possibility.

We explored three variations of PTC analyses:

1. *Single-to-single*: In our initial analyses, we perform PTC using pairs of individual (un-averaged) trials. This variation will be referred to as "single-to-single." One difficulty in performing single-to-single PTC is that when individual trials are relatively noisy or variable, as is often the case in neuroimaging data and EEG in particular, the problem is compounded by directly comparing two single recordings. Thus, performance, in this case, might be expected to be worse than typical neural network classification. By way of comparison, although traditional DNNs are trained and tested on individual trials, the network itself effectively embodies the features that worked best across the full breadth of the training set. In that sense, traditional DNNs might be a closer comparison in some ways to the single-to-average PTC analysis (see #3 below). To address the noise issue that arises when comparing two individual trials, we also performed PTC using two approaches that incorporate some form of averaging prior to classification, and use trial averages rather than individual trials as one or both elements of each input to the model.
2. *Average-to-average*: In the first averaging-based PTC variation, each training/validation/test example was composed of two trial averages that combined signals from 20 trials each, without overlap. This variation will be referred to as "average-to-average." At this point, it is important to note that most analyses are performed under the assumption that we have a pre-existing "known" dataset and a novel "unknown" dataset. As such, we explore the case in which the first signal in the pair is composed of an average generated from a single "unknown" subject, and the second signal is generated from an average across multiple "known" subjects, to form a sort of exemplar pattern. For further details, see section 2.4.
3. *Single-to-average*: In this analysis, a single trial was compared to a 20-trial average. This variation will be referred to as "single-to-average." Again, this is performed with the assumption of "known" and "unknown" datasets. The single trial is drawn from the "unknown" set, while the averaged trial is computed across multiple subjects from the "known" dataset. Again, see below for further details.

2.2.3. Dictionary Approach

The basic same/different PTC classifier can readily be mapped to multiclass classification with the use of a corpus, or *dictionary*, of "known" trials. To classify a given "unknown" sample, it is compared to sets of "known" trials from each class using PTC. The likelihood scores for each of the classes are passed through a softmax function (Bridle, 1990), and the classification decision for the unknown sample is determined by whichever dictionary class has the highest average softmaxed likelihood value.

A naïve implementation of dictionary selection was used. One hundred trials from each of the three classes were chosen at random from the test and validation set to form the dictionary. In the approaches that compare against averaged signals, each of the trials in the dictionary is created by averaging 20 randomly selected signals from the same class, chosen across subjects.

2.3. Network Architecture

Although in principle any number of architectures could be used for PTC, a straightforward choice for this initial demonstration was to use a convolutional network model, as that afforded a clear way of training a classifier that could learn to compare its two inputs. Each of the two trials is treated as an input channel, so convolution can naturally capture their parallel (both channel and time) nature[3]. A 3 × 3 2D filter was chosen to act over both channels and time, respectively, using zero-padding to maintain signal dimensionality between layers. The results were passed through a leaky rectified linear unit activation (Leaky ReLU) function and then a batch normalization layer. Four blocks of convolution, activation, and batch normalization layers were used. The number of filters per block was increased successively from 12 to 48 to capture the hierarchical nature of the feature representations. Each block was connected not only to the next block with the output of the batch normalization layer feeding into the next convolution, but all subsequent blocks with skip connections using concatenation, as in DenseNet (Huang et al., 2017). At the end of the densely connected portion, a final 2D convolution with three filters was applied to reduce the data dimensionality before feeding it into two fully-connected layers with 64 and 32 neurons, and then a final classification layer. In total, the network had 246,909 trainable parameters. While the total number of parameters is substantially smaller than found in many deep networks, previous literature has suggested that

[3] In some of our preliminary explorations of the PTC concept in another dataset, we tried combining the "trials" end-to-end with various network models, and also tried training SVMs, SMLR, and multilayer perceptrons to perform the same/different classification on similarly concatenated pairs of trials. None of these methods were able to learn to perform the classification above chance, which suggests that structuring the input to "hint" at the paired nature of the problem is necessary for an algorithm to reliably learn the same/different comparison operation. In turn, this means that DNNs, particularly convolutional architectures, may be uniquely well-suited to the task, as these architectures make it much easier to take the structure of the input into account.

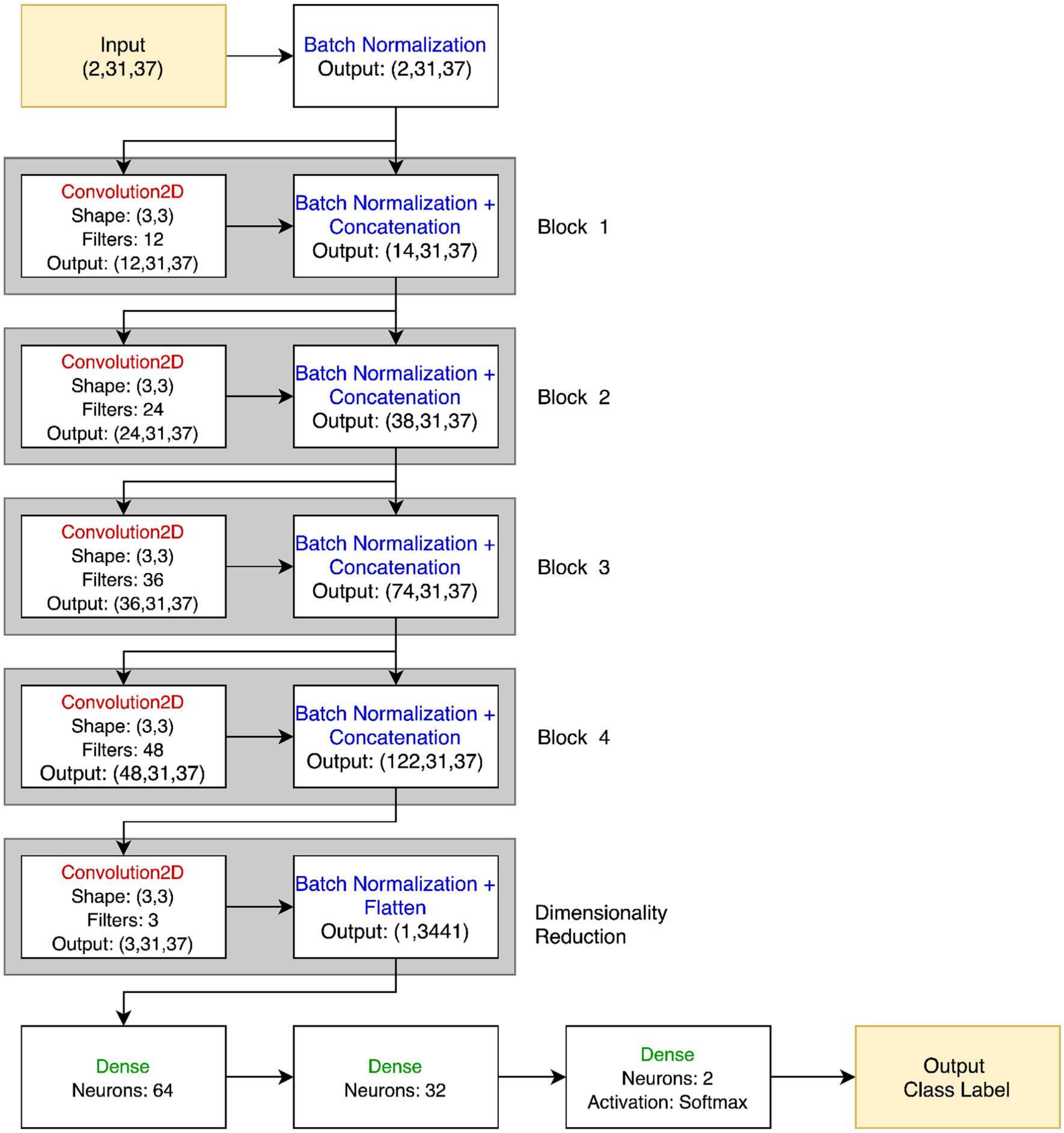


FIGURE 2 | PTC neural network architecture. The input is two 31 × 37 EEG signals, stacked together in a 2 × 31 × 37 3D array. Batch normalization is immediately applied to each signal, and the output of this is both passed through the 12 filter banks in the Convolution2D layer of Block 1 and passed directly to the Batch Normalization block. Thus, the Batch Normalization layer of Block 1 outputs 12 + 2 = 14 images of size 31 × 37. This process is repeated for a total of four blocks. No concatenation is performed in the dimensionality reduction block, and the 3 × 31 × 37 feature map is flattened and passed to the final dense layers before classification. All convolutional and dense layers use the Leaky ReLU activation function unless otherwise specified.

substantially down-scaled neural networks are appropriate for neuroimaging data, in part due to the tendency of larger networks to overfit when trained with the limited size of dataset available in neuroimaging (Bashivan et al., 2015). See **Figure 2** for a visual representation of the PTC network.

As noted above, a similar network was used for the baseline DNN model that used a more conventional three-category classification approach. The only differences were that in the baseline DNN's network, the first convolutional layer only accepted a single trial's data, and the final layer had one output per class. This version of the network thus had a very similar 245,862 trainable parameters (less than half a percent fewer than the PTC network), as the vast majority of the parameters are found in the dense layers, which share the same input and output shapes between the two models.

2.4. Training, Validation, and Test Procedures

In all the variations of PTC, subjects were split into three disjoint groups: training, validation, and test. A leave-one-subject-out cross-validation methodology was used, and the remaining subjects were split with 80% randomly assigned to training, and the remaining 20% assigned to validation. As per standard, the training group was used to perform backpropagation updates on the models; the validation group was used to determine a stopping criterion for updating the model, and the test group was used to determine the accuracy of the models.

All models were trained using the Adam optimization algorithm (Kingma and Ba, 2015) and a batch size of 144. Minibatches were generated dynamically during training. Samples were drawn randomly from all subjects in the training group, with an even split across all possible class pairs and orderings (e.g., Face-Face or Scene-Word). Since, in a three-class problem, there are six "different" pairings and three "same" pairings, the "same" pairings were sampled twice as often to provide equivalent numbers of "same" and "different" pairings during training. In the average-to-average analysis, averages were constructed such that no trials were shared between the two averages; in single-to-average, the single trial was never one of the trials used to comprise the average.

Standard techniques were implemented to reduce the likelihood of overfitting. Dropout was enforced on the dense layers (Srivastava et al., 2014), with a proportion of 10%. We initially tried architectures with higher dropout rates, which would be more standard usage of the dropout algorithm, but those rates resulted in reduced performance during training across all analyses and unreliable training convergence in the single-to-single PTC variation. Early stopping was employed when validation loss failed to improve for a period of 30 epochs. The model from the epoch in which the lowest validation loss was observed was chosen as the final model.

For single-to-single and single-to-average "same-different" accuracy, each sample in the test set (i.e., held-out subject) was compared to a randomly selected set of trials drawn from the training and validation sets, with 100 signals per class, as described in section 2.2.3. For the average-to-average PTC method, 80 averages were generated per class from the test set (roughly approximating the number of individual trials per class that a single subject would have) and then compared against a dictionary built from the training/validation sets as in the other PTC analyses.

For the baseline (non-PTC) deep learning analyses, a similar leave-one-subject-out cross-validation scheme was used, with the same network hyperparameters as the PTC analyses (optimization algorithm, dropout, early stopping, etc.). Similar to PTC, the non-left-out subjects were split with 80% randomly assigned to training, and the remaining 20% assigned to validation. For all deep-learning-based analyses, ten iterations of the cross-validation were performed per left-out subject, and results from all iterations were averaged to yield the final results we present below.

SVM and SMLR models were also tested using leave-one-subject-out cross-validation, but without a validation set. That is, the models were trained on trials from all but one subject, and the remaining subject's trials were then used for testing.

TABLE 1 | Baseline accuracy (percent, with chance = 33.33%).

	SVM	SMLR	DNN
Single	63.67	**64.90**	59.51
Averaged	81.54	82.52	**82.54**

Bold values indicate highest accuracy in each row.

2.5. Environment

All deep learning analyses (PTC analyses + the DNN baseline analysis) were performed in Python 3.6 using the Keras toolbox (Chollet, 2015) with a Theano backend (Theano Development Team, 2016). Custom in-house Python scripts were used to implement the specific analysis techniques we used, tabulate results, and so on. NumPy was used in supporting functions (Oliphant, 2006; Walt et al., 2011). SciKitLearn was used for the SVM (Buitinck et al., 2013) and PyMVPA was used for SMLR (Haxby et al., 2011).

3. RESULTS

3.1. Base Models

The performance of the three baseline classifiers is shown in **Table 1**. All values reported are derived by first calculating mean accuracies for each human participant (averaged across iterations of the cross-validation algorithm, for all deep learning models; SVM and SMLR are deterministic and did not require multiple iterations), and then averaging across participants. Overall, SMLR (with a lambda parameter of 100) achieved the highest performance on single trials, with an accuracy of 64.90% (against chance = 33.33%). SVM (with a C parameter of 0.0001) achieved the second-best results with an accuracy of 63.67%. Finally, the DNN model achieved an accuracy of 59.51%. SMLR's performance was significantly better than SVM's ($p = 0.0064$) and SVM's was significantly better than the DNN model's ($p < 10^{-7}$; all comparisons are paired *t*-tests).

In the averaged-trials condition, all baseline models performed similarly. The DNN model performed infinitesimally better than SMLR, at 82.54 and 82.52%, respectively. SVM achieved a slightly lower accuracy of 81.54%. However, none of these values were significantly different from each other (all $p > 0.3$).

3.2. PTC

The overall PTC results are shown in **Table 2**. The same-different classification had a chance accuracy of 50%, and the dictionary classification approach had a chance accuracy of 33.33%. Generally, as more averaging was applied, the accuracy increased. Same-different accuracy improved from 56.03% (single-to-single) to 71.25% (single-to-average) to 86.15% (average-to-average) as the averaging was increased. As expected, all of these values were significantly different from each other (all $p < 10^{-18}$).

Similarly, dictionary classification improved with more averaging, from 49.21 to 61.53 to 83.32%. Again, as expected, all of these values were significantly different from each other (all $p < 10^{-16}$).

TABLE 2 | PTC accuracy summary (percent).

	Same/Different	Dictionary
Single-to-Single	56.03	49.21
Single-to-Average	71.25	61.53
Average-to-Average	86.15	83.32
(Chance)	50.00	33.33

The single-to-single PTC dictionary classification performed worse than all of the single baselines ($p < 10^{-13}$ against all single-trial baseline classifiers). Similarly, the single-to-average PTC dictionary model performed worse than all of the averaged-trial baselines (all $p < 10^{-11}$). However, given the differences in the algorithms, a "fairer" comparison might be between the single-to-average PTC dictionary model and the single-trial baseline classifiers, since the trained baseline classifiers implicitly contain a form of averaged representation of the training data, against which individual trials are compared during testing. The single-to-average PTC dictionary still performed significantly worse than the single-trial SVM and SMLR classifiers (both $p \leq 0.001$), but it did outperform the single-trial DNN classifier with a nearly identical network architecture (61.53 vs. 59.51%; $p < 10^{-7}$).

The average-to-average PTC dictionary classification did perform with a numerically higher accuracy than all of the averaged-trial baselines (83.32% for PTC vs. next-highest DNN baseline at 82.54%). However, as with the comparisons among the individual averaged-trial baseline models, the average-to-average PTC dictionary classifier was not significantly different from any of them (all $p > 0.14$).

The same-different confusion matrices for the three methods are shown in **Table 3**. All three models are more likely to predict that two samples came from a different underlying class than the same underlying class, with the difference being more pronounced as more averaging is involved. As a result, accuracy was higher when the actual trial pair was a "different" pair than when it was a "same" pair.

Finally, graphical confusion matrices of the individual stimulus categories are shown for baseline classifiers and PTC analyses in **Figure 3**. Broadly speaking, all classifiers showed the same general pattern, with words being correctly identified most often, followed by scenes, followed by faces. In all cases, averaging improved performance, and various individual classifiers performed better than others, as detailed above; however, none of the classifiers or manipulations appeared to show a qualitative difference in the pattern observed in the confusion matrices, beyond those that tracked with overall increases/decreases in accuracy. As such, it appears that, broadly speaking, all classifiers were picking up on approximately the same general patterns in the data, with no classifiers or manipulations appearing to show a particular bias for one category over the others.

4. DISCUSSION

In this paper, we demonstrated a new method of deep-learning-based classification for neuroscience data, paired trial classification (PTC). Rather than using a DNN to classify a trial's category directly, we instead trained the classifier to compare pairs of trials to each other. Using a "dictionary" approach similar to ones employed in traditional MVPA studies with conventional distance/similarity metrics, we also used PTC to generate category predictions based on how often a test trial was judged to be the "same" as other trials drawn from the three categories of stimuli in our dataset. While it is difficult to draw direct performance comparisons between PTC and our baseline measures, given the significant differences in how the problem was structured and how the results could be interpreted, overall PTC performed comparably to other measures, and in some cases perhaps a bit better. Either way, we believe the novelty and flexibility of PTC make it an interesting approach and a viable avenue for future explorations into its potential.

In all cases, PTC performed with accuracy significantly above chance. While the same-different classification for the single-to-single paradigm was only marginally better than chance at 56.08% accuracy, the single-to-average paradigm was substantially better than chance with an accuracy of 71.32%. This represents the ability to say with some confidence whether a novel trial is similar to some exemplar formed from the averaging of known trials. Furthermore, the average-to-average paradigm is more accurate at 86.15% accuracy, allowing for a more confident determination of a group of novel samples known to be drawn from the same unknown class.

The tendency of the models to predict "different" more often than "same" is somewhat notable, considering the equal number of "same" and "different" examples provided during training. However, this tendency is straightforward enough to explain; it stands to reason that noise is more likely to make a trial appear as if it were coming from some different class than for noise to cause two trials from different classes to appear to be drawn from the same class. For instance, assume a research participant stops paying attention for one trial, or flutters their eyes enough to create a small artifact (but not one big enough to trigger rejection of the trial using standard preprocessing techniques). If the PTC algorithm is doing its job well, it is likely to judge that noise trial as being "different" from the trial it is paired with, regardless of whether the other trial is the same category or not. In that case, the PTC algorithm may not even be making an error when it judges some "same" pairs as "different"; instead, it might be picking up on unanticipated differences/noise in the data that are not accounted for by the comparatively simple assumption that all "face" trials, for example, should have similar neural activity to each other. This feature might be exploitable in future work; for example, to address the well-known issue with many conventional DNN analyses that deep networks often yield high confidence scores to noisy or adversarial inputs (Nguyen et al., 2015; Su et al., 2019), as implicit in their training is the tendency to maximize confidence scores as much as possible. In contrast, a PTC classifier given a poor input might correctly give high-confidence "different" responses to all of the possible categories (including the one that is nominally the same as the poor trial), which effectively can be read as a vote of no confidence in the quality of the input data.

TABLE 3 | Confusion matrices for PTC analyses (percent).

	Single-to-Single		Single-to-Average		Average-to-Average	
Predicted:	Same	Different	Same	Different	Same	Different
Actual						
Same	53.71	46.29	60.96	39.04	79.39	20.61
Different	42.74	57.26	23.50	76.50	10.47	89.53

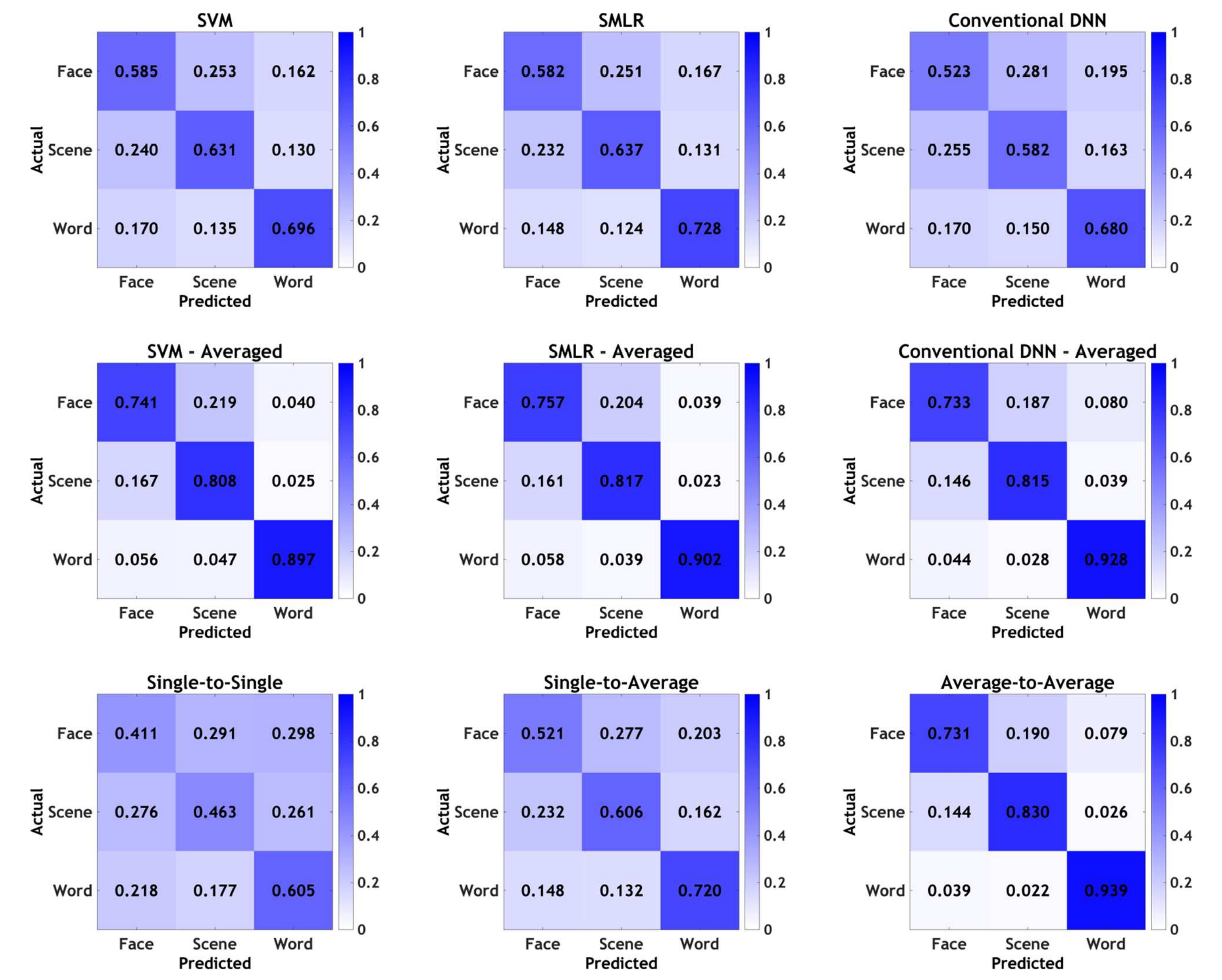


FIGURE 3 | Confusion matrices by individual stimulus categories for all classifiers. Values presented as proportions rather than percentages for readability.

We also observe that the dictionary-based classification technique allows for the successful mapping back to the multiclass classification paradigm. The results for the single-to-average dictionary classification condition were on par with any of the single trial baselines, and the average-to-average dictionary classification conditions were on par with the averaged signal baselines. These results were achieved with a naïve approach to dictionary selection, so better performance could be seen with the optimization of the dictionary; exploring potential improvements to the dictionary portion of the algorithm would be one promising direction for future work. Notably, the single-to-single paradigm stands to improve from a less noisy dictionary. As it stands in this implementation, a sort of "weak learner" effect is observed between the two applications of the single-to-single network: A single "same-different" classification was successful 56.08% of the time, only 6.08% above chance, but the three-class classification was successful 49.35% of the time, a more impressive 16.02% better than chance. Although the two values are not directly comparable given the differences in what they represent, they are suggestive that pooling the individually weaker "same-different" classifications across a multiclass dictionary can indeed produce robust overall results. This also raises the possibility that the PTC approach might be especially well-suited to datasets with higher numbers of classes. In particular, if some of those classes had too few trials in them to reliably train a conventional DNN to recognize them, a PTC network trained with trial pairs from all classes might still have a chance of picking them out.

Deep learning is useful because it can take advantage of the multi-dimensional nature of datasets in a way that other methods cannot (as the simpler linear techniques require vectorized input); GPU acceleration and parallelization in general are better supported for deep learning, making it more computationally efficient for large datasets; and deep networks can be configured flexibly to address a wider range of problem domains than simple linear methods. However, deep learning is frequently difficult to apply successfully in neuroscience. Often, human neuroscience datasets can fall into a "Goldilocks problem" zone, meaning that they can have too many trials or features for SVMs or other conventional MVPA approaches to be performant, but fewer training examples than are typically expected to enable DNN-based analyses to converge reliably. In such cases, data augmentation techniques could be applied to enable the use of deep learning by increasing the generalizability of the network. However, direct augmentation techniques pose challenges of their own. For example, Bashivan et al. (2015) found that temporal shifting techniques that have been applied successfully in other fields did not meaningfully improve generalization in their deep learning analyses of an EEG dataset. PTC offers an alternative way of increasing the size of the training dataset, not by augmenting the trial data itself, but rather by pairing trials combinatorially. However, it does not require altering the underlying data (except, optionally, by averaging trials to increase signal-to-noise, as we did here), which could be a useful property of this technique in specific scenarios, or provide another option to try when standard data augmentation techniques fail.

The goal of this paper was to introduce the PTC paradigm and show that it can easily be mapped back to multiclass classification. This approach is not limited, however, to cases in which there are a discrete, known set of classes as in typical classification applications. PTC could also be used for detecting trials that differ substantially from those seen in the training dataset, such as in outlier detection, novel stimulus identification, or artifact rejection. It could also be used in situations wherein a more conventional distance/similarity metric might be applied; for example, to assess neural pattern similarity across exposures to a set of stimuli, and to use these similarity judgments to test hypotheses about memory, make predictions about which stimulus is being seen or imagined at a particular point in time, or perform clustering analyses.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Yale University Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JW wrote the code, performed the analyses, and wrote the manuscript. AS and PR provided the guidance on methodology and co-wrote the manuscript. MJ supplied the dataset, provided the guidance on methodology, and co-wrote the manuscript.

FUNDING

This work was supported by NSF Grant CMMI 1719388, Biosensor Data Fusion for Real-time Monitoring of Global Neurophysiological Function awarded to PR and colleagues, as well as NSF/EPSCoR Grant 1632849, RII Track-2 FEC: Neural networks underlying the integration of knowledge and perception, and NIH P20 GM130461, Rural Drug Addiction Research Center, awarded to MJ and colleagues. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the University of Nebraska.

REFERENCES

Bashivan, P., Rish, I., Yeasin, M., and Codella, N. (2015). Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv* 1511.06448.

Bentin, S., Truett, A., Mccarthy, G., Puce, A., Perez, E., and McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *J. Cogn. Neurosci*. 8, 551–565. doi: 10.1162/jocn.1996.8.6.551

Bridle, J. S. (1990). "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in *Neurocomputing*, eds F. F. Soulié and J. Hérault (Berlin; Heidelberg: Springer Berlin Heidelberg), 227–236. doi: 10.1007/978-3-642-76153-9_28

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., et al. (2013). "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* (Würzburg), 108–122.

Chollet, F. (2015). *Keras*. Available online at: https://keras.io

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn*. 20, 273–297.

Haxby, J., Guntupalli, J., Connolly, A., Halchenko, Y., Conroy, B., Gobbini, M., et al. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* 72, 404–416. doi: 10.1016/j.neuron.2011.08.026

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 4700–4708. doi: 10.1109/CVPR.2017.243

Inoue, H. (2018). Data augmentation by pairing samples for images classification. *arXiv* 1801.02929.

Johnson, M. R., McCarthy, G., Muller, K. A., Brudner, S. N., and Johnson, M. K. (2015). Electrophysiological correlates of refreshing: event-related potentials associated with directing reflective attention to face, scene, or word representations. *J. Cogn. Neurosci.* 27, 1823–1839. doi: 10.1162/jocn_a_00823

Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *International Conference on Learning Representations 2015* (San Diego, CA), 1–15.

Krishnapuram, B., Carin, L., Figueiredo, A. T., Member, S., and Hartemink, A. J. (2005). Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 957–968. doi: 10.1109/TPAMI.2005.127

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., and Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *J. Neural Eng.* 15:056013. doi: 10.1088/1741-2552/aace8c

Li, W., Wu, G., Zhang, F., and Du, Q. (2016). Hyperspectral image classification using deep pixel-pair features. *IEEE Trans. Geosci. Rem. Sens.* 55, 844–853. doi: 10.1109/TGRS.2016.2616355

Lotte, F., Bougrain, L., Cichocki, A., Clerc, M., Congedo, M., Rakotomamonjy, A., et al. (2018). A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *J. Neural Eng.* 15:031005. doi: 10.1088/1741-2552/aab2f2

Nguyen, A., Yosinski, J., and Clune, J. (2015). "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 427–436. doi: 10.1109/CVPR.2015.7298640

Oliphant, T. (2006). *A Guide to NumPy*. Trelgol Publishing. Available online at: https://www.scipy.org/citing.html

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp*. 38, 5391–5420. doi: 10.1002/hbm.23730

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. doi: 10.5555/2627435.2670313

Su, J., Vargas, D. V., and Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput*. 23, 828–841. doi: 10.1109/TEVC.2019.2890858

Theano Development Team (2016). Theano: a Python framework for fast computation of mathematical expressions. *arXiv* abs/1605.02688.

Theiler, J. (2013). Matched-pair machine learning. *Technometrics* 55, 536–547. doi: 10.1080/00401706.2013.838191

Valenzi, S., Islam, T., Jurica, P., and Cichocki, A. (2014). Individual classification of emotions using EEG. *J. Biomed. Sci. Eng*. 7, 604–620. doi: 10.4236/jbise.2014.78061

Walt, S. V. D., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng*. 13, 22–30. doi: 10.1109/MCSE.2011.37

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 Neuroscience

ORIGINAL RESEARCH
published: 21 May 2020
doi: 10.3389/fnins.2020.00260

A 3D Fully Convolutional Neural Network With Top-Down Attention-Guided Refinement for Accurate and Robust Automatic Segmentation of Amygdala and Its Subnuclei

Yilin Liu[1], *Brendon M. Nacewicz*[2], *Gengyan Zhao*[3], *Nagesh Adluru*[1], *Gregory R. Kirk*[1], *Peter A. Ferrazzano*[1,4], *Martin A. Styner*[5,6] and *Andrew L. Alexander*[1,2,3]*

[1] Waisman Brain Imaging Laboratory, University of Wisconsin-Madison, Madison, WI, United States, [2] Department of Psychiatry, University of Wisconsin-Madison, Madison, WI, United States, [3] Department of Medical Physics, University of Wisconsin-Madison, Madison, WI, United States, [4] Department of Pediatrics, University of Wisconsin-Madison, Madison, WI, United States, [5] Department of Psychiatry, University of North Carolina-Chapel Hill, Chapel Hill, NC, United States, [6] Department of Computer Science, University of North Carolina-Chapel Hill, Chapel Hill, NC, United States

OPEN ACCESS

Edited by:
Yogesh Rathi,
Harvard Medical School,
United States

Reviewed by:
Roberto Viviani,
University of Innsbruck, Austria
Hans J. Johnson,
The University of Iowa, United States

***Correspondence:**
Andrew L. Alexander
alalexander2@wisc.edu

Specialty section:
This article was submitted to
Brain Imaging Methods,
a section of the journal
Frontiers in Neuroscience

Received: *13 October 2019*
Accepted: *09 March 2020*
Published: *21 May 2020*

Citation:
Liu Y, Nacewicz BM, Zhao G, Adluru N, Kirk GR, Ferrazzano PA, Styner MA and Alexander AL (2020) A 3D Fully Convolutional Neural Network With Top-Down Attention-Guided Refinement for Accurate and Robust Automatic Segmentation of Amygdala and Its Subnuclei.
Front. Neurosci. 14:260.
doi: 10.3389/fnins.2020.00260

Recent advances in deep learning have improved the segmentation accuracy of subcortical brain structures, which would be useful in neuroimaging studies of many neurological disorders. However, most existing deep learning based approaches in neuroimaging do not investigate the specific difficulties that exist in segmenting extremely small but important brain regions such as the subnuclei of the amygdala. To tackle this challenging task, we developed a dual-branch dilated residual 3D fully convolutional network with parallel convolutions to extract more global context and alleviate the class imbalance issue by maintaining a small receptive field that is just the size of the regions of interest (ROIs). We also conduct multi-scale feature fusion in both parallel and series to compensate the potential information loss during convolutions, which has been shown to be important for small objects. The serial feature fusion enabled by residual connections is further enhanced by a proposed top-down attention-guided refinement unit, where the high-resolution low-level spatial details are selectively integrated to complement the high-level but coarse semantic information, enriching the final feature representations. As a result, the segmentations resulting from our method are more accurate both volumetrically and morphologically, compared with other deep learning based approaches. To the best of our knowledge, this work is the first deep learning-based approach that targets the subregions of the amygdala. We also demonstrated the feasibility of using a cycle-consistent generative adversarial network (CycleGAN) to harmonize multi-site MRI data, and show that our method generalizes well to challenging traumatic brain injury (TBI) datasets collected from multiple centers. This appears to be a promising strategy for image segmentation for multiple site studies and increased morphological variability from significant brain pathology.

Keywords: deep learning, fully convolutional neural network, amygdala, structural MRI, segmentation, harmonization, generalization

1. INTRODUCTION

The amygdala is a key regulator of emotional arousal and is thought to regulate generalization or habituation of fear responses in normal and abnormal development (Adolphs et al., 2005; Knight et al., 2005; Öhman, 2005). Animal models have been used to differentiate subregions of the amygdala, identifying structural bases of fear generalization in basal and lateral nuclei distinct from output projections from centromedial regions (Amaral et al., 1992; LeDoux, 2007; Hrybouski et al., 2016; Kwapis et al., 2017), and reliable quantification of these substructures would be extremely useful. Accurate segmentation of the amygdala and specific subregions for quantitative analyses may provide better insights into fear and emotion processing and the role of the amygdala in traumatic brain injury and neuropsychiatric diseases. However, as a deep heterogeneous cluster of subregions, surrounded by vasculature, it remains an extremely difficult region to quantify. Compared with conventional automated software (Freesurfer, FSL), hand drawn amygdala boundaries can better capture cumulative contributions of biological and environmental stress, including autistic social impairment, physical abuse, institutional neglect and poverty (Nacewicz et al., 2006; Hanson et al., 2015). However, manual segmentation is extremely time-consuming and is prone to biases (Maltbie et al., 2012), highlighting the need for highly accurate automated segmentation methods. Currently, there are no reliable segmentation tools for subnuclei regions of the amygdala. Furthermore, the effects of image and subject variability from scanner, protocol and brain pathology on amygdala segmentation have not been previously investigated.

Segmentation methods for the amygdala can largely be classified into atlas-based and learning-based categories. A high resolution MRI atlas of the amygdala with defined subregions was recently described (Tyszka and Pauli, 2016); however, the utilization of this atlas to individual brain images is limited by the ability to anatomically spatially align the atlas. A promising strategy is the multi-atlas based method in which the segmentation of a target image is estimated by aligning it with one or more labeled atlases through registration (Babalola et al., 2009; Leung et al., 2010; Hanson et al., 2012). There is, however, a considerable computational cost associated with multi-atlas approaches since all of the atlases need to be deformably registered to each target image case using non-linear deformable transformations (Hanson et al., 2012). Additionally, the segmentation quality in multi-atlas approaches highly depends on the selection of the atlases and the fusion algorithm (Rohlfing et al., 2004; Aljabar et al., 2009). Other automatic population atlas-based segmentation packages are FreeSurfer and FSL, but overall their segmentation performances remain not optimal (Morey et al., 2009; Schoemaker et al., 2016) due to insensitivity to biologically-relevant variance (Hanson et al., 2015) and failure to capture subtle boundaries of centromedial nuclei when applied to single subjects (Saygin et al., 2017). Furthermore, neither Freesufer nor FSL support the segmentation of the subregions of the amygdala.Therefore, neither Freesurfer nor FSL performance are evaluated in this paper. A significant limitation with existing tools and prior work in this domain is that the effects of variability across scanners and protocols have not been investigated, nor have the effects of brain injuries on amygdala segmentation.

Recently, convolutional neural networks (CNN) have brought tremendous improvements in various computer vision tasks such as image classification and segmentation (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; He et al., 2016). Unlike traditional machine learning, CNN as a learning based approach can autonomously learn representations of data with increasing levels of abstraction via multiple convolutional layers without feature engineering. In CNNs, weights are shared and locally connected among convolutional layers, which significantly reduces the number of parameters compared with fully connected layers, making CNNs especially suitable for imaging tasks. Naturally, CNNs have been gradually becoming the tool of choice for medical imaging tasks. In medical image segmentation, a classification network was previously proposed using a sliding window scheme to predict the class probability of the center pixels of over-lapping patches (Ciresan et al., 2011). Since such a classification makes predictions for a single pixel at a time, this approach suffers from redundant computations and does not benefit from correlations across pixels. Long et al. (2015) first proposed then fully convolutional neural networks (FCNN) in which the fully connected layers are replaced with 1x1 convolution so that the network consists of convolutional layers only. This strategy allows dense predictions for multiple pixels in a single forward pass, and eliminates the limitation posed by fully connected layers on the size of the input image size. FCNN therefore serves as an effective general purpose engine for tasks of semantic image segmentation.

A widely-used FCNN architecture is "encoder-decoder," which are popularized by U-Net (Ronneberger et al., 2015), 3D U-Net (Çiçek et al., 2016), V-Net (Milletari et al., 2016), and SegNet (Badrinarayanan et al., 2017). The encoder part compresses the input images into lower-resolution feature maps via downsampling or pooling layers, and the decoder part aims to recover the full-resolution label map from these feature maps for pixel-to-pixel semantic classification. These networks have similar encoders—a VGG-like (Simonyan and Zisserman, 2014) architecture is typically adopted, while they vary with respect to their decoder strategies. Multiple up-sampling strategies have been proposed for decoders, including deconvolution (Noh et al., 2015), bilinear upsampling and unpooling (Badrinarayanan et al., 2017). However, such design could pose a few problems when segmenting structures with small spatial extent. First, although consecutive strided convolutions or pooling operations employed in these networks enable a large receptive field, fine details may be lost and are difficult to remedy via simple non-learnable upsampling strategies or skip connections. For example, if a network has a downsample rate of 1/8 (as it employs three max-pooling layers with 2 × 2 filters with stride 2), an object with less than 8 voxels (such as the amygdala's subregions) in each dimension may not be well recovered later. Second, since down-sampling operations typically lead to great dimension reduction, the input images of these networks need to be large enough so as to preserve sufficient dimension after the compression of the encoder, for being further processed by the decoder. But larger image patches are more likely to be dominated by background voxels compared with smaller ones, leading to severe class imbalance problem. This makes the predictions more favorable to the background, which is particularly of concern for small objects. Although a weighted cross entropy loss function has been suggested to alleviate this problem (Ronneberger et al., 2015; Çiçek et al., 2016), choosing a proper weight map for all the classes is non-trivial. Another solution could be the Dice loss function (Milletari et al., 2016) which avoids tuning any extra hyperparameter and weighs false negatives and false positives equally. Hence, although these networks have plenty of success in segmentation tasks of large structures such as brain extraction (Zhao et al., 2018), lung (Negahdar et al., 2018), and breast segmentation (Dalmış, 2017), specific strategies for small structures are necessary.

Compared with larger structures, smaller ones like the amygdala and its subregions provide fewer signals to exploit, which makes the learning of discriminative features more challenging. Hu and Ramanan (2017) suggested that modeling context is particularly helpful for CNNs to recognize small objects, based on a key observation that humans can only accurately classify small faces with evidence beyond the object itself. In general, context can provide knowledge of a structure with respect to its surroundings and disambiguate objects with similar local visual appearances. Thus, incorporating context can critically improve recognition accuracy (Galleguillos and Belongie, 2010). In medical imaging, many studies have explored the idea of using input patches with various sizes for modeling multi-scale contextual information (de Brebisson and Montana, 2015; Moeskops et al., 2016; Ghafoorian et al., 2017; Kamnitsas et al., 2017). Most of these networks are organized in a multi-branch manner, where each branch independently processes patches of a certain type. In other patch-based CNN approaches, explicit spatial features obtained from a structural probabilistic atlas are combined with CNN features to provide additional spatial information (Kushibar et al., 2018). Another line of efforts focuses on enlarging kernels via dilated convolutions to integrate larger contextual information (Chen et al., 2018). Segmenting small structures with high accuracy is therefore reduced to the problem of finding the optimal trade-off between capturing sufficiently large context and retaining fine details, while alleviating the imbalanced class issue.

In light of the limitations of previous works, we present a dual-branch dilated residual FCNN with two parallel convolutions to extract both local context for alleviating the class imbalance issue and more global context. Residual connections (He et al., 2016) are added to facilitate the gradient flow and more importantly, feature reuse from earlier layers. In order to enhance such feature fusion, we additionally develop a top-down attention-guided (AG) refinement unit resided on residual connections to select useful low-level details from earlier layers to better complement the highly semantic feature maps from deep layers, which we believe can benefit the segmentation of small regions like the amygdala and subnuclei on structural T1-weighted images. In general, attention mechanisms can emphasize important features and suppress the irrelevant ones, mimicking human visual system, which has been broadly applied to various vision and natural language processing tasks (Bahdanau et al., 2014). A popular attention mechanism, "Squeeze & Excitation" (SE) module (Hu et al., 2018) which recalibrates channels by modeling channel interdependencies, has been shown to be effective in some medical images segmentation tasks (Roy et al., 2018). Different from SE, we utilize higher-level information as priors to recalibrate lower-level channels.

This study focused on two critical areas of brain image segmentation—(1) the parcellation of very small structures like the subnuclei of the amygdala, and (2) the application of whole amygdala segmentation across multiple scanners and variable brain injuries. For the parcellation of amygdala subnuclei, we evaluated the accuracy of our segmentation method by comparing it to other automated methods including two deep learning based and a multi-atlas based method. A preliminary version of the presented work appeared in Liu et al. (2018). We further demonstrate the benefits of the dual-branch design by analyzing the influence of each branch on final performance and compare the two design choices of our attention-guided refinement unit to SE module (Hu et al., 2018), showing that the top-down AG refinement unit is more suitable than SE in this application, and potentially in segmentation tasks of other small structures. Finally, we investigated a strategy to generalize the FCNN amygdala segmentation approach to a challenging Traumatic Brain Injury (TBI) dataset collected from multiple sites, despite the variability of contrast and image sensitivity across MRI scanner hardware (RF coils, in particular) and software (pulse sequences and protocols) and increased image heterogeneity associated with pathology, demonstrating its robustness to real-world practice.

2. MATERIALS AND METHODS

2.1. Dataset

T1-weighted MRI data from 14 subjects (age mean (standard deviation) 28.9 years (6.5 years); range 18.5–43.4 years), each imaged in both morning and evening sessions on 2 days separated by 1 week (four total imaging sessions) on a GE MR750 3.0 T MRI scanner with the product 8-channel head coil. All participants provided written consent or assent as part of a procedure approved by the Human Subjects Institutional Review Board of the University of Wisconsin School of Medicine and Public Health. A whole-brain 3D inversion-recovery prepared fast gradient-echo T1- weighted sequence (inversion time TI = 600 ms; fast gradient echo readout TR/TE = 9.4/3.1 ms; 256 × 192 matrix, resampled to 256 × 256, over 240 mm field of view with 128 slices 1-mm thick) was prescribed as axial oblique slices angled so that the midpoint and splenium of the corpus callosum occupied the same plane (Nacewicz et al., 2012).

An iterative pre-processing pipeline used 3DSkullStrip (AFNI) (Cox, 1996) to output a roughly skull-stripped image, which was then coregistered to the MNI152 template by affine transform in FLIRT (FSL) (Jenkinson et al., 2012), and tissue priors reverse warped to native space for segmentation-based bias-field correction in FAST (FSL), the dilated bias field was applied to the original image, which was then more effectively skull-stripped, contrast-adjusted and squared to exaggerate gray matter-CSF differences, re-coregistered to the MNI template for better alignment of tissue priors and a final bias correction with FAST. This method was developed to preserve tissue in the lateral nucleus of the amygdala, which is otherwise frequently misclassified as CSF and erroneously darkened by bias corrections. The resultant images from each of the 4 sessions were coregistered in FLIRT to an individual-subject averaged space (1 mm isotropic) representing an affine transform equidistant to all 4 session images and averaged (Nacewicz et al., 2012), followed by landmark-based AC-PC alignment with concomitant cropping to 191 × 236 × 171 (Cox, 1996), and rotation to the "pathological plane" to match post-mortem atlases (Nacewicz et al., 2006).

Both the left and the right amygdala were manually divided on the 4-session averaged 1 mm isotropic T1-weighted images into four subnuclear groups on each side—lateral, basal, cortico-superficial (olfactory) and centromedial subregions—by an amygdala anatomy expert (BN) based on visible landmarks largely matching those described by Amaral et al. (1992). Details of how subregions are defined are provided in **Figure 1**. We note that the slight in-plane downsampling to 1mm and the spatial normalization did not impair the manual labeling. Specifically, in the coronal plane the *lateral* nucleus was easily isolated due to its darker intensity. The combined *basal* nuclei went from the thin white matter capsule of the cortical nucleus medially to the intense, linear lateral border formed by the fibers passing through the plane along this edge and with careful effort to include the magnocellular "dogleg" portion and its white matter capsule; and the dorsomedial boundary of the basomedial region was formed by a straight line from the most ventromedial extent of the visible white matter around the "dogleg" down to the most ventromedial tip of the amygdala clearly visible above the hippocampal head. The combined *cortico-superficial* nuclei included all tissue bordering on the ambient cistern above the semiannular sulcus ventrally up to the more lateral of either the rhinal sulcus or lateral extent of the optic tract, with the lateral boundary defined by the white matter capsule of the cortical nucleus or the straight line boundary described above for the basal group. The combined *centromedial* group was bounded by white matter dorsally including a thin boundary between the central nucleus and putamen, extended ventromedially along the white matter forming the dorsomedial boundary of the "dogleg" of the basolateral nucleus, then a straight line extended dorsomedially to the more lateral of the rhinal sulcus or optic tract. The manual labeling of 10 ROIs per individual on 14 brains with two blinded repeats (four amygdalae) yielded intra-rater Dice overlap coefficients: Lateral = 0.89, Basal = 0.82, Centromedial = 0.77, Superficial = 0.75 and total amygdala using our previously published technique yielded excellent agreement (dice = 0.94). Manual tracing is, however, quite tedious and time-intensive, requiring 10–20 person-hours per brain, which limits application to larger data sets. Overall, the right and the left amygdala jointly account for about 0.05% of the whole brain volume of a single subject. Training and evaluation of the segmentation methods as described below were performed on single session (non-averaged) data using the segmentation labeling from the averaged data.

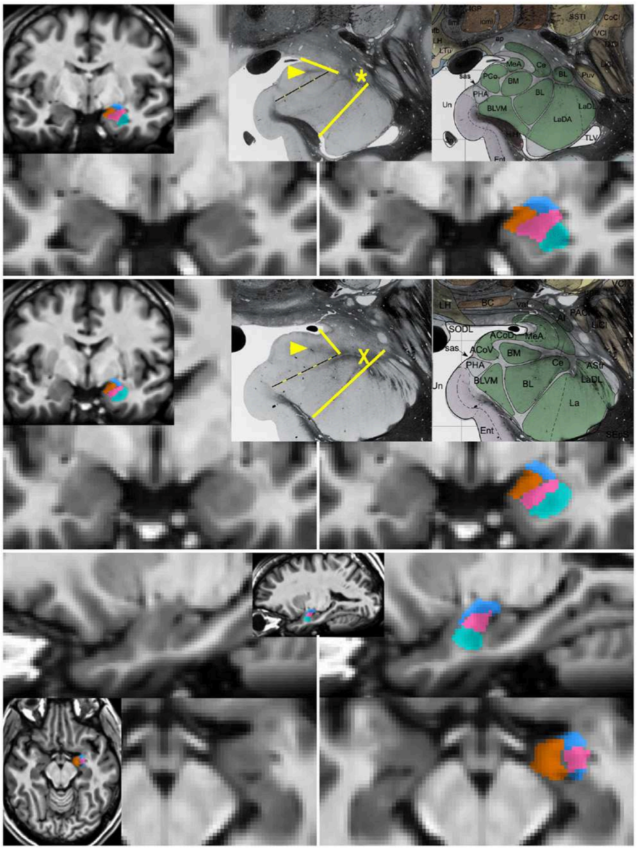


FIGURE 1 | Segmentation of subnuclear groups by landmarks visible on single subject images. Unlabeled (left) and labeled (right) images at more posterior (top) and anterior (middle) coronal sections with representative histology and subdivisions from Mai et al. (3rd ed) (Mai et al., 2015). Tracing began in the coronal section with the "dogleg" of the basolateral nucleus (asterisk). The lateral nucleus (teal) was easily identifiable by the lower T1 intensity lateral to a linear border with the basolateral nucleus. The combined basal nuclei (pink) was defined starting in the plane of the dogleg, with the dorsal boundary following the thin white matter angling inferomedially along the central nucleus. The medial boundary of the basal group extends up to but not including the white matter encircling the cortical nucleus. A key landmark anterior to the dogleg is a spider-like white matter formation (middle, X) dividing all subdivisions and discernible in all single-subject images. When the white matter of the cortical nucleus was not visible, a spot of white matter at the triple junction with the medial nucleus (arrowhead in top and middle) or the most medial tip of white matter between basolateral and central nucleus was connected with the most medial extent of the subventricular/uncal white matter (dotted line). The cortico-superficial grouping (orange) extends superiorly to a line from the triple junction in posterior sections or the tip of white matter above basolateral nucleus on anterior sections to the more superolateral of the endorhinal sulcus or optic tract. The centromedial group (blue), includes all darker tissue above these boundaries. All nuclei were then refined to achieve smooth agreement in sagittal and axial views (bottom).

2.2. Network Backbone

To incorporate larger contexts while alleviating class imbalance, we present a dual-branch model design (**Figure 2**), with one specializing in capturing multi-scale contexts and the other maintaining a small receptive field which helps the model focus on the ROIs. For any given feature map $U \in \mathbb{R}^{H\times W\times D}$, kernels of two different sizes are applied in parallel to perform two transformations $\Omega : U \rightarrow \hat{F} \in \mathbb{R}^{H' \times W' \times D'}$ and $\psi : U \rightarrow \tilde{F} \in \mathbb{R}^{H' \times W' \times D'}$, forming two branches. In order to more efficiently preserve information, dilated convolutions (Yu and Koltun, 2015) in place of down-sampling layers are adopted throughout the network, i.e., kernels are up-sampled with zeros inserted between weights so that the receptive field of the kernels can be expanded without incurring extra computational costs. The gap between elements in a kernel is $D_k - 1$, where D_k denotes the dilation rate, with standard convolution as a special case when D_k is 1. Therefore, the two branches are composed of 3^3 kernels with $D_{k_1} \geq 1$ (dilated branch) and $D_{k_2} = 1$ (standard branch), respectively. For example, a 5^3 kernel for the dilated branch is a 3^3 kernel with $D_k = 2$. Batch Normalization (Ioffe and Szegedy, 2015) and ReLU non-linearity (Glorot et al., 2011) are applied in sequence after convolutions. Information from both branches are then fused via element-wise summations before being fed into the next layer (**Figure 3**, left):

$$F^l = \hat{F}^l_{dilated} + \tilde{F}^l_{normal},$$

where F^l denotes the fused feature maps (FMs) for each layer l. The small dilation rates designed for standard branch are to ensure that it has a small receptive field of size 19 × 19 × 19 which can just enclose the whole amygdala. This allows for a detailed analysis of the ROIs and alleviates the class imbalance problem, since the receptive field determines the number of voxels that can influence model predictions per optimization step. For the dilated branch, the dilation rates are empirically set to be $D_{k_1} = \{1, 2, 4, 2, 8, 2, 4, 2, 1\}$, resulting in a receptive field of size 53 × 53 × 53, which can capture large contexts. The number of kernels for each branch is as follows: 30, 30, 40, 40, 40, 40, 50, 50, 50. In addition to such parallel feature fusion, residual connections (He et al., 2016) are also integrated into the network mainly for feature reuse (Chen et al., 2017) in series, which adds the features from a lower layer to those from a higher layer via skip connections (**Figure 3**, right). Both the parallel and serial feature fusion are shown in **Figure 3**. They are further enhanced by a top-down attention mechanism described in section 2.3.

2.3. Top-Down Attention-Guided Refinement Unit

CNNs are known to have an inherent feature hierarchy, where layers that are close to the inputs extract high-resolution spatial details and deeper layers form highly semantic but coarser

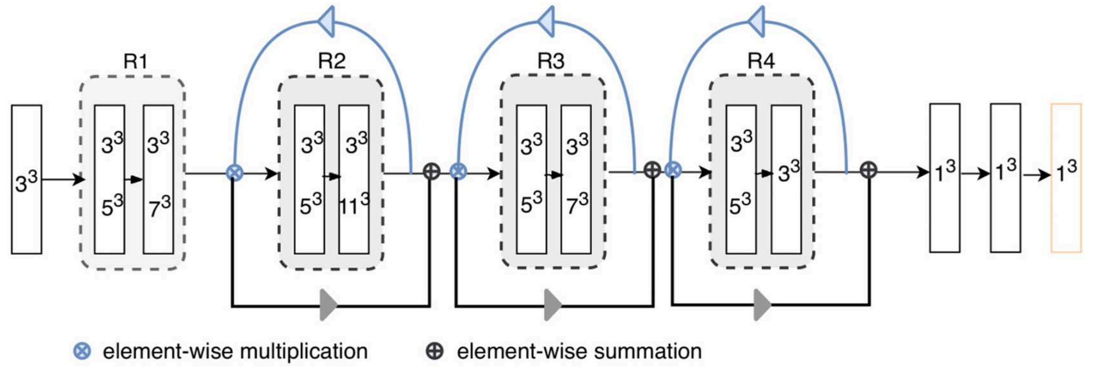


FIGURE 2 | Architecture of the proposed model. "RX"s represent residual blocks (the residual connections are omitted here). The rectangles with two kernel sizes represent parallel convolutions, as illustrated in **Figure 3**. The attention weights generated using higher-level feature priors, denoted as blue arrows, are multiplied with the lower-level channels; then, the reweighted lower-level features are used to refine the next layers, as shown by gray arrow. Each layer except for the final classification layer (orange) is followed by batch normalization and ReLU.

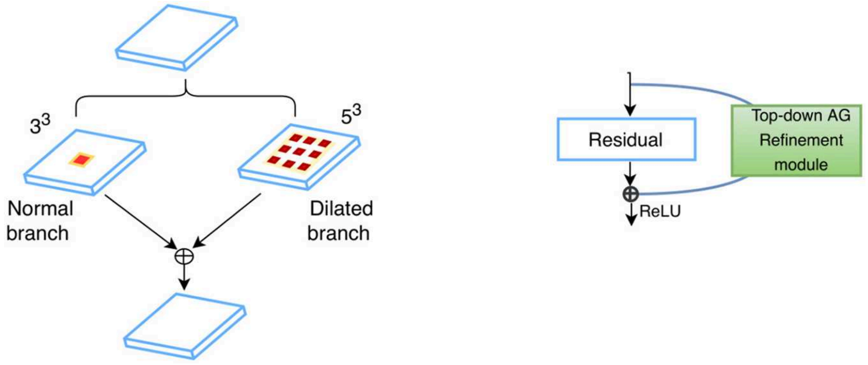


FIGURE 3 | Feature fusion in parallel **(Left)** and series **(Right)**: kernels of two different sizes are applied in parallel, and the resultant feature maps are fused via element-wise summation; standard residual connections are adopted for serial feature fusion, where features from earlier layers are incorporated into deeper layers.

features. A number of deep learning studies have explored to fuse multi-level features from different layers to enrich the feature representation (Hariharan et al., 2015; Long et al., 2015; Ronneberger et al., 2015; Lin et al., 2017; Zhang et al., 2018). Especially, segmentation of small objects is found to benefit from such feature reuse from earlier layers where fine-grained low-level details are abundant (Shrivastava et al., 2016; Lin et al., 2017). Nevertheless, indiscriminately fusing the different levels of features may not always be effective due to the semantic dissimilarity empirically found by Zhang et al. (2018). Motivated by their observation, we propose a top-down attention-guided refinement unit based on residual connections to supplement the typical feed-forward, bottom-up CNN, where the abundant semantic information from the higher layers can highlight and select the low-level details from lower layers, as shown in **Figure 4**. Given a set of features maps from earlier layers $F_{low} \in \mathbb{R}^{C' \times H \times W \times D}$, a set from higher layers $F_{high} \in \mathbb{R}^{C'' \times H \times W \times D}$, and the attention coefficients $\alpha \in \mathbb{R}^{1 \times 1 \times 1 \times C'}$ the *refined* feature maps from higher layers can be defined as:

$$F'_{high} = F_{high} + d(\alpha \otimes F_{low}),$$

where $\otimes$ denotes element-wise multiplication, $F = F_{dilated} + F_{normal}$ for all layers,and $d(\cdot)$ represent $1 \times 1 \times 1$ convolutions for aligning the dimensionality of that of the higher-level feature maps. α is formulated as the following:

$$\alpha = [\alpha_1, \alpha_2, ..., \alpha_c],$$

$$\alpha_c = \sigma(Z(B(Conv^{1\times1\times1}(AvgPool(F_{high}))))),$$

where Z represents the rectified linear unit (ReLU) function, which provides non-linearity by setting negative values as zeros and keeping positive ones constant; B denotes the batch normalization (Ioffe and Szegedy, 2015) , which can accelerate and stabilize network training by standardizing each training batch; and σ denotes the sigmoid function for rescaling the attention coefficients to $[0, 1]$.

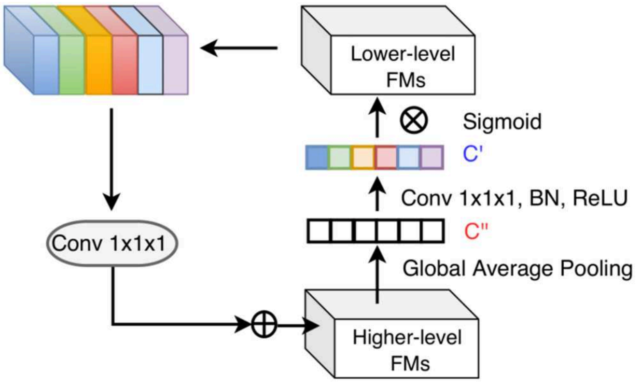


FIGURE 4 | Top-down attention-guided refinement unit on residual connections, where lower-level features are recalibrated by higher-level information and incorporated into deeper layers. “FMs” denotes as feature maps. Channel-wise statistics of higher-level information are first extracted by global average pooling, and the interdependencies among channels are modeled by a $1 \times 1 \times 1$ convolution followed by the sigmoid activation. The reweighted lower-level features are then added to the higher-level features.

2.4. Evaluation in a Multi-Site Data Set With Brain Pathology

Amygdala segmentation strategies with CNN methods were also evaluated in a T1-weighted structural imaging study of children ages 9–18 years with severe traumatic brain injury (TBI) scanned 1–2 years after the injury. Twenty-one children (13F/8M) ages 9–18 years were scanned with T1w imaging at 13 sites with differing 3T MRI scanner systems, RF coils and pulse sequences. Among the TBI scans, 9 sites scanned one subject, 3 sites scanned two subjects and 1 site scanned six subjects. Representative images are shown in **Figure 5**. The data collection was approved by the Institutional Review Boards for each site and parental assent and informed consent was obtained for all subjects. Similar imaging protocols were employed across sites (3D T1w MP-RAGE (TI = 900 ms on Siemens and Philips) or BRAVO IR-fSPGR (TI = 450 ms on GE) with 1 mm isotropic spatial resolution (256 mm FOV with 256 × 256 matrix and 192 sagittal slices at 1 mm thick); however, there was variability between sites in terms of scanner manufacturers and models, RF coils, and pulse sequences, which affected spatial sensitivity, contrast, and image quality. Further, the severity, type and localization of injuries was extremely heterogeneous across sites. All these issues pose challenges on the applicability of CNNs, which typically do not generalize well to data whose distribution is different from that of the training data (Gibson et al., 2018a). Prior studies on multi-site generalized segmentation either retrains the model directly on multi-site data (Gibson et al., 2018a) or fine-tunes the domain-specific

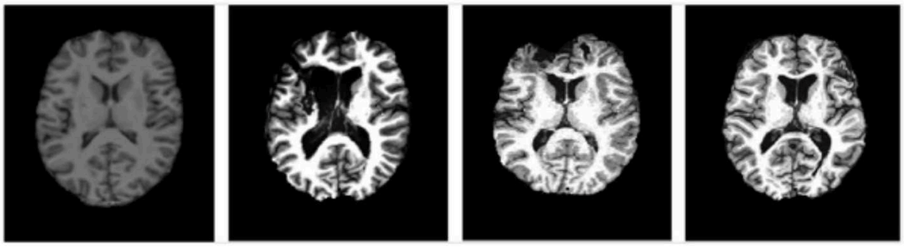


FIGURE 5 | Representative images at similar anatomic levels from the source domain (a healthy subject, the leftmost) and target domains (3 TBI patients in the 3 rightmost frames). The slices were selected to highlight the lesion pathology and not the amygdala.

parameters (Karani et al., 2018) of the model, both requiring a few labeled target images from the new sites. In this study, we instead resort to pixel-level image adaptation, aiming to directly segment the full amygdala volumes from the multi-site images without the corresponding labels. We did not attempt to evaluate the segmentation of amygdala subregions for this multi-site study because manual labeling was deemed impractical for these data due to insufficient data quality for reliable identification.

As there was considerable site-to-site variability, we investigated the utility of a cycle-consistent generative adversarial network approach (CycleGAN) (Zhu et al., 2017) to harmonize the image contrast with the training data. CycleGAN has not been applied to multi-site data harmonization before, to the best of our knowledge. Specifically, the distribution of multi-site target data is transformed into source-like distribution while the appearance of the target images are preserved. In this way, a pre-trained segmentation model can be directly applied to the adapted target images without prior assumptions on scanner/protocol deviations. CycleGAN consists of two generators that learn two mappings, respectively, $G_1 : S \rightarrow T$ and $G_2 : T \rightarrow S$, and two discriminators D_1, D_2 that distinguish the generated images from the real ones for each domain. In particular, we are interested in the generator G_2 that transforms the target images into realistic source-like images, i.e., $G_2(x^t) = x^{t\rightarrow s}$. The distribution of the target and source images are aligned by applying adversarial losses (Goodfellow et al., 2014) where G tries to confuse D by producing realistic source-like images. Cycle-consistent losses (Zhu et al., 2017) computed by l_1 distance are also applied to ensure that the generated target images are similar to the original ones. The transformed target images eventually obtained from the CycleGAN will be rendered as if they are drawn from the source domain, with the contents preserved. The total loss is defined as:

$$\mathcal{L}_{total}(G_1, G_2, D_1, D_2) = \mathcal{L}_{adv}(G_1, D_2) + \mathcal{L}_{adv}(G_2, D_1) \\ + \lambda \mathcal{L}_{cyc}(G_1, G_2),$$

where λ is used to modulate the strength of the cycle consistency.

2.5. Implementation Details

The proposed segmentation method was implemented in PyTorch, using one Titan Xp GPU for training. Categorical cross entropy was employed as the cost function, optimized via the Adam solver with an initial learning rate of 0.001, scheduled to decay as $lr = lr_{initial} * \left(1 - \frac{iter_n}{total_{iter}}\right)^{power}$, where *power* was set to 0.9. Weights in each layer were initially drawn from a zero-based Gaussian distribution with standard deviation of $\sqrt{2/n_i}$, where n denotes the number of units in a kernel of the layer l (He et al., 2016). Bias were initialized at zero. Training was performed in batches of 14 image patches. In each iteration, 11 patches of size $59 \times 59 \times 59$ were sampled from the whole brain and fed into the model. During inference, $105 \times 105 \times 105$ patches were used. For comparison, training of the other deep learning based methods, i.e., *HighRes3DNet* (Li et al., 2017), *DeepMedic* (Kamnitsas et al., 2017) were implemented in Tensorflow (Gibson et al., 2018b) following their original settings in the respective papers, i.e., Dice loss (Milletari et al., 2016) was used in *HighRes3DNet* and categorical cross entropy in *DeepMedic*. An existing multi-atlas based method (Wang et al., 2014) was also evaluated for comparison in a leave-one-out fashion: 13 atlases were used for training and one atlas for evaluation. For all the deep learning based methods evaluated, a 7-fold cross validation was performed. In each fold, 10 subjects were used for training, 2 for validation and 2 for testing. The models were trained with a fixed number of epochs. The model parameters in the epoch that resulted in best performance (i.e., highest average dice) on the validation set were used to segment the test set. Performance of all methods on the test set was reported.

For multi-site MR image harmonization, we trained the CycleGAN on the coronal view of all the images from all domains. For the architecture choices, we followed the original settings: two convolutions with stride of 2, 9 residual blocks, two fractionally strided convolutions with stride $\frac{1}{2}$ are employed as the generator (Johnson et al., 2016), and 70×70 PatchGAN (Isola et al., 2017) is employed as the discriminator which aims to detect 70×70 image patches as real or fake. In total 3,304 slices from the source data and 5,900 slices from the TBI data are used for training. Each slice is randomly cropped to 128×128 before being fed into the CycleGAN. Data augmentation includes random rotation with angles of $\gamma \cdot 90^\circ$, where $\gamma \in [0, 1, 2, 3]$, and scaling with factors $0.8, 1, 1.2$. For comparison only, we also conducted supervised training by training a model using the labeled TBI data in a 7-fold cross validation scheme, and the above-mentioned multi-atlas based method which was trained on the source data in a leave-one-out cross validation scheme and then directly applied to the TBI data. Results are summarized and analyzed in section 3.4.

2.6. Evaluation Metrics

The pair-wise similarity and discrepancy of our automatic (A) and manual segmentation (M) were evaluated using the commonly employed Dice Similarity Coefficient (DSC):

$$DSC = \frac{2|A \cap M|}{|A| + |M|},$$

whose value ranges from zero to 1, where 1 indicates 100% with the ground truth, and 0 indicates no overlap. However, volumetric overlap measures are not sensitive to the contour of the segmentation output, while the latter is important in many medical applications such as disease diagnosis and treatment planning, as is also the case for the amygdala (Shenton et al., 2002; Tang et al., 2015; Yoon et al., 2016). Thus, we additionally consider a distance-based metric—the average symmetric surface distance (ASSD) (Geremia et al., 2011) in our evaluation. ASSD is defined as the average of distances between border voxels of our automatic segmentation output and those of manual segmentation output:

$$ASSD \\ = \frac{\sum_{m \in B(M)} min_{a \in B(A)} ||m - a|| + \sum_{a \in B(A)} min_{m \in B(M)} ||a - m||}{|B(M)| + |B(A)|},$$

where $B(\cdot)$ denotes the set containing all the voxels on the border. Zero value for this measure indicates a perfect segmentation.

3. RESULTS

In this section, we present qualitative and quantitative results for our model and conduct ablation studies to demonstrate the effectiveness of each proposed component. We also compare the results of the proposed method with several state-of-the-art methods on the same dataset. Finally, we explore the feasibility of harmonizing the multi-site TBI data using CycleGAN and show the generalized capability of our method. Wilcoxon signed rank tests (two-sided) are used for performance comparison throughout the analysis.

3.1. Single-Branch vs. Dual-Branch

Here we demonstrate the advantages of the dual-branch design by investigating the influences of each single branch. Experiments of using the dilated and standard branch separately are conducted in the same 7-fold cross validation scheme. Each branch is equipped with residual connections as in the original dual-branch setting. It can be observed in **Table 1** that the *dilated* branch, which has a significantly larger receptive field, performs better on larger subregions (lateral, basal), while the *standard* branch with a smaller receptive field is better at segmenting smaller subregions, especially on the cortico-superficial subregions ($p = 0.007$). Additionally, the *dilated* branch yields significantly lower ASSD values than the *standard* branch on all subregions ($p<0.05$). The *dual-branch* network inherits the merits of each single branch and achieves best overall accuracy in terms of both Dice and ASSD. Qualitative results of the compared models are shown in **Figure 6**.

3.2. Top-Down Attention-Guided Refinement Unit

We also tested the effectiveness of the proposed top-down attention guided feature refinement scheme for further boosting the accuracy. Two variants were explored: "local reweighting" and "global reweighting," as illustrated in **Figure 7**. These were compare with the SE blocks (Hu et al., 2018) that are also placed on the residual connections. **Table 2** shows that the "local reweighting" scheme yields best overall Dice, especially on the cortical-superifical subregions ($p < 0.05$) which are the most challenging due to the smallest volume-to-surface ratio. Thus, we employ a "local reweighting" scheme for the attention module. Meanwhile, we can observe that the addition of either the "global reweighting" scheme or the SE blocks results in comparable or increased model complexity, while the results get slightly worse. This demonstrates that the improvements are indeed due to better feature refinement resulting from the locally top-down attention module, and not simply from the increased capacity of the model.

3.3. Comparison With Other State-of-the-Art Methods

In order to demonstrate the advantage of the proposed method, we compared our method with some other popular publicly available segmentation methods including two deep learning

TABLE 1 | Dice overlap (columns 2–4) and ASSD (columns 5–7) performance of both single branch models and the dual-branch model.

Subregions		Dice (%)			ASSD (mm)	
	Dilated	standard	Dual	Dilated	standard	Dual
Lateral	80.6 (6.6)	77.9 (7.7)	**82.6 (5.0)**	0.70 (0.24)	2.66 (1.90)	**0.68 (0.31)**
Basal	76.6 (6.6)	75.9 (6.1)	**77.3 (6.0)**	**0.70 (0.15)**	1.10 (0.68)	0.71 (0.20)
Centromedial	73.7 (7.7)	**76.7 (5.2)**	75.4 (5.3)	**0.61 (0.16)**	1.00 (0.66)	0.61 (0.20)
Cortical-Superficial	71.7 (5.7)	72.2 (5.6)	**73.1 (5.6)**	0.96 (0.44)	1.94 (2.00)	**0.81 (0.33)**
Mean	75.6 (7.4)	75.7 (6.5)	**77.1 (6.4)**	0.74 (0.30)	1.67 (1.59)	**0.70 (0.27)**

Subregions are listed in descending order by their volume-to-surface ratio. Highest are highlighted in bold and the second highest are underlined. The dual-branch model performance was either highest or second highest for all regions in terms of both Dice overlap or ASSD.

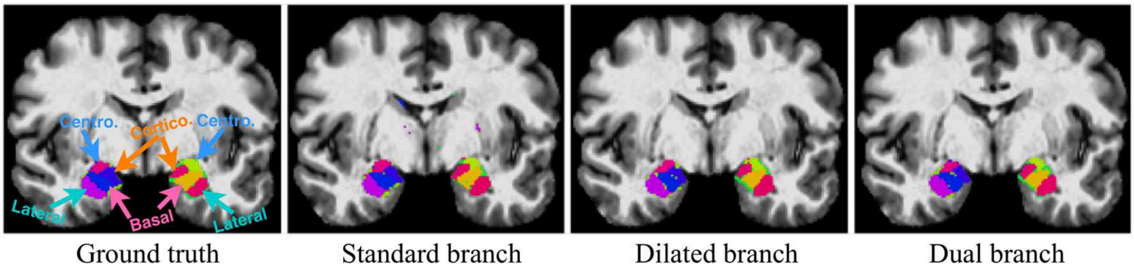


FIGURE 6 | Qualitative segmentation examples show influences of each single branch on the final dual-branch model. The incorporation of larger context (Dilated branch) enables the final model to better localize the subregions, thus reducing false positives (the scattered misclassified background voxels, as seen on the Standard Branch result), while standard branch helps refine the appearance details of the final output.

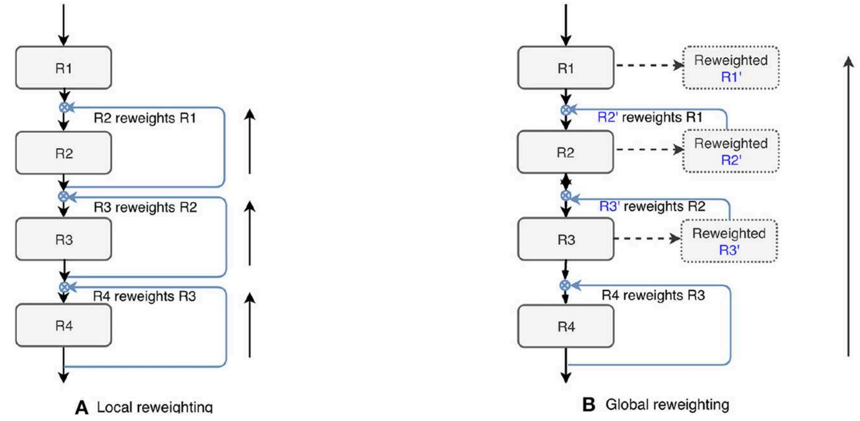


FIGURE 7 | Two variants of the proposed top-down attention. RX denotes the residual blocks (residual connections are omitted here).

TABLE 2 | Comparison for the Dice score (%) of the two variants and the SE blocks against the baseline (dual-branch model) and the percentage increase in model complexity.

Subregions	Baseline	*SE*	Global	Local
Lateral	82.6 (5.0)	81.2 (7.1)	**83.4 (5.1)**	82.8 (5.2)
Basal	77.3 (6.0)	76.9 (5.7)	77.2 (5.5)	**77.6 (5.3)**
Centromedial	75.4 (5.3)	74.5 (6.2)	76.3 (5.1)	**76.6 (5.6)**
Cortical-Superficial	73.1 (5.6)	71.7 (5.1)	72.5 (5.8)	**74.7 (5.6)**
Mean	77.1 (6.4)	76.1 (6.9)	77.4 (6.7)	**78.0 (6.1)**
Parameters (% increase)	0.795M (–)	0.811M (+2.0%)	0.808M (+1.6%)	0.808M (+1.6%)

The largest value in each row is bold faced.

models, *DeepMedic* and *HighRes3DNet*, and a multi-atlas based algorithm. *HighRes3DNet* is a state-of-the-art method in brain parcellation for 155 neuroanatomical structures (not including extremely small brain structures such as the subregions of the amygdala), and *DeepMedic* has shown excellent performance in lesion segmentation. Results (**Table 3**) show that our method exhibited superior performance in terms of both Dice and ASSD in this application. The differences in Dice with *DeepMedic* on the lateral ($p = 0.04$), basal ($p = 0.03$) and cortical-superficial ($p < 0.005$) subregions were significant. In particular, our method demonstrated substantial improvements for the cortical-superficial subregions thanks to the top-down attention guided refinement module. *DeepMedic* performed better ASSD on the basal subregions ($p < 0.005$) and our method were better at the cortical-superficial subregions ($p < 0.03$). Compared to multi-atlas, our method yielded significantly better Dice on the lateral, basal and cortical-superficial subregions ($p < 0.05$; $p < 0.05$; $p < 10^{-3}$, respectively). There was no statistically significant differences on ASSD between our method and the multi-atlas based method.

3.4. Generalization on Multi-Site TBI Dataset

Whole-amygdala segmentation performance on the training data is reported in **Table 4**, which shows a roughly 90% overlap between the algorithm and ground truth. We investigated the generalization of the proposed method on a challenging multi-site TBI dataset by directly applying the trained whole-amygdala segmentation model to the TBI data. The results were evaluated relative to the "gold standard" defined by manual correction of Freesurfer amygdala segmentations by an expert (GK). Both Dice overlaps and ASSD were computed. For comparison only, we also conducted supervised training with TBI labels (corrected Freesurfer segmentations). As the objective was to evaluate the utility of CycleGAN for improving deep neural network (DNN)'s performance when testing on out-of-distribution data, the performance of competing CNN methods on the multi-site TBI data was not evaluated for these data. It is clear from **Table 5** that a direct application of our trained model to the multi-site data demonstrated very poor performance, while after harmonization by CycleGAN, the trained model's performance on target data was significantly improved ($p < 10^{-6}$). Supervised training yielded slightly higher performance. The multi-atlas based method, which is much less affected by the shift in data distribution, demonstrated similar performance to our method after harmonization, though the processing time is considerably longer. It should be noted that the segmentation performance for all the approaches was substantially lower than for the segmentation applied to the training data (**Table 4**). Qualitative results for one subject are shown in **Figure 8**.

4. DISCUSSION

In this study, we present a lightweight dual-branch residual FCNN with enhanced feature refinement to segment the subregions of the amygdala. Parallel branches with different dilation rates are used to process objects with different scales as well as extract more global contexts, and a top-down attention-guided refinement unit is proposed to guide the selection of lower level details for better feature refinement. We evaluated our method on MRI image data acquired from a cohort of adolescents. The results show that the proposed method achieved better performance as compared to several existing state-of-the-art segmentation methods. Meanwhile, our approach takes several seconds to segment the data of a subject, which is orders of magnitude faster than the multi-atlas based approach. This opens up the potential for real-time use during MRI acquisition, which would facilitate individualized functional, structural or spectroscopic imaging of small anatomical structures.

From the results of using each branch separately, we found that the performance on objects of different scales can be critically influenced by the receptive fields, and the proper receptive fields is correlated with the scale of objects. Our dual-branch design with different receptive fields thus flexibly adapt to subregions of different scales. Furthermore, although the standard branch with a small receptive field is prone to spatial inconsistencies due to local similarities, dilated branch remedies this effect by incorporating more global contextual information via dilated convolutions. The significantly lower ASSD values it yields suggest that dilated branch is especially effective in reducing such false positives, indicating its strong localization ability for ROIs and boundaries. This suggests that each branch provides complementary information toward the solution of the segmentation problem. Benefiting from both branches, the final model obtained substantially more accurate segmentation results both volumetrically and morphologically.

Besides the lightweight dual-branch backbone, we also explore the idea of multi-scale fusion and enhance it with a top-down attention-guided refinement unit. An important design choice for the proposed refinement unit is the strategy to use more local or global high-level information as the guide. The results indicate that the local refinement scheme may be more suitable and it is especially advantageous in small and challenging subregions (cortical-superficial). This is consistent with our hypothesis that smaller objects tend to benefit more from feature reuse. Interestingly, the comparison with SE blocks suggest that SE blocks inhibit rather than emphasize the ROIs in this application, as also found in Roy et al. (2018). This may due to the small size of the features of the ROIs whose contribution to the whole feature maps are less significant compared with other features of the same level and are thus suppressed. We therefore speculate that the top-down design, which utilizes higher semantic and categorical information as priors to determine the importance the lower-level features, may alleviate this problem and thus may be more suitable for segmentation tasks of small objects.

In comparisons with two other state-of-the-art deep learning models, our method shows superior performance in terms of both Dice overlap and ASSD. Notably, all evaluated models contain comparable parameters and therefore comparable capacities, while they vary in their topological structures. HighRes3DNet consists of consecutive 20 dilated residual convolutional layers with progressively enlarged dilation rates. It shares many key components with the backbone of our model such as the dilated residual convolutions, but has them connected in series only while ours also in parallel. Such serial connections result in an overly large receptive field (87 × 87 × 87) which causes severe class imbalance in segmenting small and compact subregions that cannot seem to be well resolved by using Dice loss, as indicated in **Table 3**. This also demonstrates the benefit of having an another branch that maintains a small receptive field in our model design. DeepMedic consists of two independent branches with the second branch processing a low-resolution version of the inputs. Compared with HighRes3DNet, the architecture of DeepMedic is flexible enough to process input segments with smaller spatial sizes, which can inherently balance the distribution of different classes. DeepMedic also exploits multi-scale learning scheme, but the responses of two branches are not fused until the very end of

TABLE 4 | Dice overlap performance on the main training dataset using a leave-one-out approach (described in section 2.1).

Amygdala	**L. Amyg**	**R.Amyg**	**Mean**
Dice (%)	90.6 (2.1)	90.5 (2.1)	90.6 (1.9)

This trained model is also applied to the harmonized TBI dataset.

TABLE 3 | Mean and standard deviation of the Dice scores and ASSD for the proposed method, two other state-of-the-art deep learning based and a multi-atlas based segmentation methods evaluated on subregions.

Methods	**Lateral**	**Basal**	**Centromedial**	**Cortical-Superficial**	**Mean**
DICE (%)					
Multi-atlas	80.3 (7.0)	75.4 (6.1)	75.2 (6.4)	69.9 (5.7)	75.2 (7.3)
HighRes3DNet	68.1 (11.4)	69.3 (7.0)	25.3 (34.5)	65.8 (6.7)	57.1 (26.1)
DeepMedic	80.5 (7.5)	75.6 (6.5)	75.5 (5.3)	71.6 (4.2)	75.8 (6.7)
Dual (Ours)	82.6 (5.2)	77.3 (6.0)	75.4 (5.3)	73.1 (5.6)	77.1 (6.4)
Dual + Top-down Att (Ours)	**82.8 (5.0)**	**77.6 (5.3)**	**76.6 (5.7)**	**74.7 (5.4)**	**78.0 (6.1)**
ASSD (mm)					
Multi-atlas	**0.60 (0.20)**	0.73 (0.16)	**0.54 (0.12)**	0.75 (0.16)	**0.66 (0.18)**
HighRes3DNet	2.00 (1.26)	1.20 (0.43)	16.63 (12.20)	1.18 (0.51)	5.25 (8.96)
DeepMedic	1.13 (1.11)	**0.52 (0.36)**	0.76 (0.67)	1.37 (1.01)	0.94 (0.89)
Dual (Ours)	0.67 (0.31)	0.71 (0.20)	0.61 (0.20)	0.81 (0.33)	0.70 (0.27)
Dual + Top-down Att (Ours)	0.94 (1.30)	0.69 (0.15)	0.67 (0.42)	**0.73 (0.22)**	0.76 (0.70)

"Dual" denotes the proposed segmentation model without the top-down attention guided feature refinement module. Highest are highlighted in bold and the second highest are underlined.

TABLE 5 | Performance before and after harmonization using CycleGAN and supervised training using TBI labeled data, and a multi-atlas method.

Settings	No harmonization	After harmonization	Supervised	Multi-atlas
Dice (%)	42.4 (21.8)	75.5 (6.7)	76.0 (9.6)	75.0 (8.4)
ASSD (mm)	N/A	1.2 (0.7)	1.9 (1.7)	0.9 (2.9)

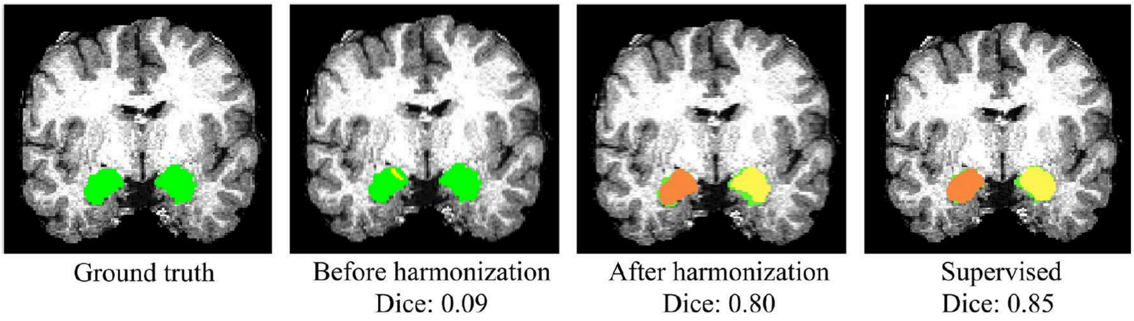


FIGURE 8 | Qualitative results of whole-amygdala segmentation in a single TBI scan. Automated segmentation results are shown in orange and yellow, and the ground truth expert labeled segmentations are shown in green. The overlays show that the segmentation was very poor before CycleGAN harmonization (2nd column), but much improved after harmonization.

the network. In contrast, our model encourages interactions of multi-resolution features both in parallel and in series. This could explain the improved performance of even our dual-branch model with respect to DeepMedic, though they have the same model complexity.

Finally, we evaluate the generalizability of our method on a multi-site TBI dataset by first pre-training the model on the main dataset and then directly applying it to the TBI data. In order to address domain shifts, we explore the feasibility of harmonizing the multi-site data using CycleGAN, which is shown to be effective and nearly closes the gap to supervised training (i.e., training with TBI labels) in this application. Comparing the Dice overlap performance of the supervised training on the main dataset and the TBI dataset, the accuracy drop on TBI data (90% to 76%) may be attributed to high variations due to heterogeneous scanning methods and anatomical injuries. Thus, larger labeled datasets are desired for better training for TBI studies, which however are often not feasible in medical imaging where expert-defined labels are often rare. Our results show that after a decent data harmonization by CycleGAN, using a single small set ($N \approx 14$) of high-quality labeled data (even though they are healthy subjects) can approximate the accuracy of directly training with a few ($N \approx 21$) TBI labeled data. This suggests that our solution makes it possible to reuse labels from different domains and thus alleviate the burdens for labeling. Another important advantage is that knowledge of sources of biases from scanners/protocols are not required for harmonization using CycleGAN. A limitation, however, is that CycleGAN only adapts images at pixel-level while feature spaces should ideally be aligned as well for better domain adaptation, which we leave for future works. Another limitation with this study was that only the whole amygdala segmentations were evaluated because the raw T1-weighted images were not of sufficient quality for expert manual labeling of the subregions.

5. CONCLUSION

In this study, we presented a novel dual-branch dilated residual FCNN with enhanced feature fusion via a top-down attention-guided refinement unit to segment the subregions of the amygdala with high accuracy. Each branch with a different receptive field demonstrated specialized ability of processing objects of the corresponding scale, thus providing complementary information. Also, we found that the proposed attention-guided feature refinement module may be more suitable than the SE blocks in segmenting small structures due to the top-down design. The proposed model showed superior performance compared with two state-of-the-art deep learning methods. Our method also shows decent generalizability on a challenging multi-site TBI dataset without needing to be re-trained, after harmonizing the TBI data using a CycleGAN. We believe that our findings and the model design could provide insights especially on generalized segmentation of small objects, which are relatively under-studied, and the high efficiency of our technique will potentially benefit real-time use in clinical practices.

DATA AVAILABILITY STATEMENT

The datasets generated/analyzed for this study can be found in https://www.nitrc.org/projects/amyg_autoseg.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Wisconsin - Madison Health Sciences IRB. Written informed consent to participate in this study was provided by the participant, or the participants' legal guardian/next of kin).

AUTHOR CONTRIBUTIONS

YL, GZ, NA, MS, and AA contributed to the conception and design of the work. BN, PF, and AA contributed to the acquisition of the data for the study. YL, BN, MS, GK, and GZ contributed to the analyses and interpretation of the data. YL wrote the first draft of the manuscript. BN, MS, and AA provided significant contributions to the writing. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

This work was supported by NARSAD: Brain and Behavior grant 24103 (to BN) and National Institutes of Health grant funding NINDS R01 NS092870, NIMH P50 MH100031 and the Waisman Center U54 IDDRC from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (U54 HD090256). NA was also supported in part by the BRAIN Initiative R01-EB022883-01, the Center for Predictive and Computational Phenotyping (CPCP) U54-AI117924-03, the Alzheimer's Disease Connectome Project (ADCP) UF1-AG051216-01A1 and R56-AG052698-01.

REFERENCES

Adolphs, R., Gosselin, F., Buchanan, T. W., Tranel, D., Schyns, P., and Damasio, A. R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature* 433:68. doi: 10.1038/nature03086

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *NeuroImage* 46:726–38. doi: 10.1016/j.neuroimage.2009.02.018

Amaral, D. G., Price, J. L., Pitkanen, A., and Carmichael, S. T. (1992). "Anatomical organization of the primate amygdaloid complex," in *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction*, ed J. P. Aggleton (New York, NY: Wiley-sLiss), 1–66. Available online at: https://search.library.wisc.edu/catalog/999682167302121

Babalola, K. O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., et al. (2009). An evaluation of four automatic methods of segmenting the subcortical structures in the brain. *Neuroimage* 47:1435–1447. doi: 10.1016/j.neuroimage.2009.05.029

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.* 39:2481–2495. doi: 10.1109/TPAMI.2016.2644615

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv: 1409.0473.*

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

Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., and Feng, J. (2017). "Dual path networks," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 4467–4475.

Çiçek, O., 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. doi: 10.1007/978-3-319-46723-8_49

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. (2011). "Flexible, high performance convolutional neural networks for image classification," in *Twenty-Second International Joint Conference on Artificial Intelligence* (Barcelona).

Cox, R. W. (1996). Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res*. 29, 162–173. doi: 10.1006/cbmr.1996.0014

Dalmış, M. U., Litjens, G., Holland, K., Setio, A., Mann, R., Karssemeijer, N., and Gubern-Mérida, A. (2017). Using deep learning to segment breast and fibroglandular tissue in MRI volumes. *Med. Phys*. 44, 533–546. doi: 10.1002/mp.12079

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

Galleguillos, C., and Belongie, S. (2010). Context based object categorization: a critical survey. *Comput. Vision Image Understand*. 114, 712–722. doi: 10.1016/j.cviu.2010.02.004

Geremia, E., Clatz, O., Menze, B. H., Konukoglu, E., Criminisi, A., and Ayache, N. (2011). Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *NeuroImage* 57, 378–390. doi: 10.1016/j.neuroimage.2011.03.080

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I. W. M., Sanchez, C. I., Litjens, G., et al. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Sci. Rep*. 7, 1–12. doi: 10.1038/s41598-017-05300-5

Gibson, E., Hu, Y., Ghavami, N., Ahmed, H. U., Moore, C., Emberton, M., et al. (2018a). "Inter-site variability in prostate segmentation accuracy using deep learning," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018* (Granada: Springer International Publishing), 506–14. doi: 10.1007/978-3-030-00937-3_58

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., et al. (2018b). Niftynet: a deep-learning platform for medical imaging. *Comput. Methods Progr. Biomed*. 158, 113–122. doi: 10.1016/j.cmpb.2018.01.025

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., 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* (Montreal, QC), 2672–2680.

Hanson, J., Suh, J., Nacewicz, B., Sutterer, M., Cayo, A., Stodola, D., et al. (2012). Robust automated amygdala segmentation via multi-atlas diffeomorphic registration. *Front. Neurosci*. 6:166. doi: 10.3389/fnins.2012.00166

Hanson, J. L., Nacewicz, B. M., Sutterer, M. J., Cayo, A. A., Schaefer, S. M., Rudolph, K. D., et al. (2015). Behavioral problems after early life stress: contributions of the hippocampus and amygdala. *Biol. Psychiatry* 77, 314–323. doi: 10.1016/j.biopsych.2014.04.020

Hariharan, B., Arbelaez, P., Girshick, R., and Malik, J. (2015). "Hypercolumns for object segmentation and fine-grained localization," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA: IEEE). doi: 10.1109/CVPR.2015.7298642

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: IEEE), 770–778. doi: 10.1109/CVPR.2016.90

Hrybouski, S., Aghamohammadi-Sereshki, A., Madan, C. R., Shafer, A. T., Baron, C. A., Seres, P., et al. (2016). Amygdala subnuclei response and connectivity during emotional processing. *Neuroimage* 133, 98–110. doi: 10.1016/j.neuroimage.2016.02.056

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE). doi: 10.1109/CVPR.2018.00745

Hu, P., and Ramanan, D. (2017). "Finding tiny faces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 951–959. doi: 10.1109/CVPR.2017.166

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

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1125–1134. doi: 10.1109/CVPR.2017.632

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision* (Amsterdam: Springer), 694–711. doi: 10.1007/978-3-319-46475-6_43

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

Karani, N., Chaitanya, K., Baumgartner, C., and Konukoglu, E. (2018). "A lifelong learning approach to brain MR segmentation across scanners and protocols," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018* (Granada: Springer International Publishing), 476–484. doi: 10.1007/978-3-030-00928-1_54

Knight, D. C., Nguyen, H. T., and Bandettini, P. A. (2005). The role of the human amygdala in the production of conditioned fear responses. *Neuroimage* 26, 1193–1200. doi: 10.1016/j.neuroimage.2005.03.020

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.

Kushibar, K., Valverde, S., González-Villà, S., Bernal, J., Cabezas, M., Oliver, A., et al. (2018). Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Med. Image Anal*. 48, 177–186. doi: 10.1016/j.media.2018.06.006

Kwapis, J. L., Alaghband, Y., López, A. J., White, A. O., Campbell, R. R., Dang, R. T., et al. (2017). Context and auditory fear are differentially regulated by hdac3 activity in the lateral and basal subnuclei of the amygdala. *Neuropsychopharmacology* 42:1284. doi: 10.1038/npp.2016.274

LeDoux, J. (2007). The amygdala. *Curr. Biol*. 17, R868–R874. doi: 10.1016/j.cub.2007.08.005

Leung, K. K., Barnes, J., Ridgway, G. R., Bartlett, J. W., Clarkson, M. J., Macdonald, K., et al. (2010). Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *Neuroimage* 51, 1345–1359. doi: 10.1016/j.neuroimage.2010.03.018

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M. J., and Vercauteren, T. (2017). "On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task," in *Lecture Notes in Computer Science*, eds M. Niethammer, M. Styner, S. Aylward, H. Zhu, I. Oguz, P.-T. Yap, D. Shen (Boone, NC: Houston, TX: Springer International Publishing), 348–360. doi: 10.1007/978-3-319-59050-9_28

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE), 2117–2125. doi: 10.1109/CVPR.2017.106

Liu, Y., Nacewicz, B., Kirk, G., Alexander, A., and Adluru, N. (2018). "Cascaded 3d fully convolutional neural network for segmenting amygdala and its subnuclei," in *Proceedings of the Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM)* (Paris).

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. doi: 10.1109/CVPR.2015.7298965

Mai, J. K., Majtanik, M., and Paxinos, G. (2015). *Atlas of the Human Brain*. Cambridge, UK: Academic Press.

Maltbie, E., Bhatt, K., Paniagua, B., Smith, R. G., Graves, M. M., Mosconi, M. W., et al. (2012). Asymmetric bias in user guided segmentations of brain structures. *Neuroimage* 59, 1315–1323. doi: 10.1016/j.neuroimage.2011.08.025

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)* (California, CA: IEEE), 565–571. doi: 10.1109/3DV.2016.79

Moeskops, P., Viergever, M. A., Mendrik, A. M., de Vries, L. S., Benders, M. J. N. L., and Isgum, I. (2016). Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans. Med. Imaging* 35, 1252–1261. doi: 10.1109/TMI.2016.2548501

Morey, R. A., Petty, C. M., Xu, Y., Hayes, J. P., Wagner II, H. R., Lewis, D. V., et al. (2009). A comparison of automated segmentation and manual tracing for quantifying hippocampal and amygdala volumes. *Neuroimage* 45, 855–866. doi: 10.1016/j.neuroimage.2008.12.033

Nacewicz, B. M., Angelos, L., Dalton, K. M., Fischer, R., Anderle, M. J., Alexander, A. L., et al. (2012). Reliable non-invasive measurement of human neurochemistry using proton spectroscopy with an anatomically defined amygdala-specific voxel. *Neuroimage* 59, 2548–2559. doi: 10.1016/j.neuroimage.2011.08.090

Nacewicz, B. M., Dalton, K. M., Johnstone, T., Long, M. T., McAuliff, E. M., Oakes, T. R., et al. (2006). Amygdala volume and nonverbal social impairment in adolescent and adult males with autism. *Arch. Gen. Psychiatry* 63, 1417–1428. doi: 10.1001/archpsyc.63.12.1417

Negahdar, M., Beymer, D., and Syeda-Mahmood, T. F. (2018). "Automated volumetric lung segmentation of thoracic CT images using fully convolutional neural network," in *Medical Imaging 2018: Computer-Aided Diagnosis*, eds K. Mori and N. Petrick (SPIE). doi: 10.1117/12.2293723

Noh, H., Hong, S., and Han, B. (2015). "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago: IEEE). doi: 10.1109/ICCV.2015.178

Öhman, A. (2005). The role of the amygdala in human fear: automatic detection of threat. *Psychoneuroendocrinology* 30, 953–958. doi: 10.1016/j.psyneuen.2005.03.019

Rohlfing, T., Brandt, R., Menzel, R., and Maurer, C. R. Jr. (2004). Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage* 21, 1428–1442. doi: 10.1016/j.neuroimage.2003.11.010

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

Roy, A. G., Navab, N., and Wachinger, C. (2018). "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018* (Granada: Springer International Publishing), 421–429. doi: 10.1007/978-3-030-00928-1_48

Saygin, Z. M., Kliemann, D., Iglesias, J. E., van der Kouwe, A. J., Boyd, E., Reuter, M., et al. (2017). High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: manual segmentation to automatic atlas. *Neuroimage* 155, 370–382. doi: 10.1016/j.neuroimage.2017.04.046

Schoemaker, D., Buss, C., Head, K., Sandman, C. A., Davis, E. P., Chakravarty, M. M., et al. (2016). Hippocampus and amygdala volumes from magnetic resonance images in children: assessing accuracy of FreeSurfer and FSL against manual segmentation. *Neuroimage* 129, 1–14. doi: 10.1016/j.neuroimage.2016.01.038

Shenton, M. E., Gerig, G., McCarley, R. W., Székely, G., and Kikinis, R. (2002). Amygdala–hippocampal shape differences in schizophrenia: the application of 3d shape models to volumetric MR data. *Psychiatry Res*. 115, 15–35. doi: 10.1016/S0925-4927(02)00025-2

Shrivastava, A., Sukthankar, R., Malik, J., and Gupta, A. (2016). Beyond skip connections: top-down modulation for object detection. *arXiv:1612.06851*.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*.

Tang, X., Holland, D., Dale, A. M., Younes, L., Miller, M. I., and Initiative, A. D. N. (2015). The diffeomorphometry of regional shape change rates and its relevance to cognitive deterioration in mild cognitive impairment and Alzheimer's disease. *Hum. Brain Mapp*. 36, 2093–2117. doi: 10.1002/hbm.22758

Tyszka, J. M., and Pauli, W. M. (2016). *In vivo* delineation of subdivisions of the human amygdaloid complex in a high-resolution group template:

in vivo amygdala subdivisions. *Hum. Brain Map.* 37, 3979–3998. doi: 10.1002/hbm.23289

Wang, J., Vachet, C., Rumple, A., Gouttard, S., Ouziel, C., Perrot, E., et al. (2014). Multi-atlas segmentation of subcortical brain structures via the autoseg software pipeline. *Front. Neuroinformatics* 8:7. doi: 10.3389/fninf.2014.00007

Yoon, S., Kim, J. E., Kim, G. H., Kang, H. J., Kim, B. R., Jeon, S., et al. (2016). Subregional shape alterations in the amygdala in patients with panic disorder. *PLoS ONE* 11:e0157856. doi: 10.1371/journal.pone.0157856

Yu, F., and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*.

Zhang, Z., Zhang, X., Peng, C., Xue, X., and Sun, J. (2018). "ExFuse: enhancing feature fusion for semantic segmentation," in *Computer Vision–ECCV 2018* (Munich: Springer International Publishing), 273–288. doi: 10.1007/978-3-030-01249-6_17

Zhao, G., Liu, F., Oler, J. A., Meyerand, M. E., Kalin, N. H., and Birn, R. M. (2018). Bayesian convolutional neural network based MRI brain extraction on nonhuman primates. *NeuroImage* 175, 32–44. doi: 10.1016/j.neuroimage.2018.03.065

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE). doi: 10.1109/ICCV.2017.244

Conflict of Interest: AA is Co-owner of Thervoyant, Inc.

The remaining 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 Neuroscience

ORIGINAL RESEARCH
published: 18 June 2020
doi: 10.3389/fnins.2020.00557

A Deep Learning-Based Model for Classification of Different Subtypes of Subcortical Vascular Cognitive Impairment With FLAIR

Qi Chen[1]†, Yao Wang[2]†, Yage Qiu[2], Xiaowei Wu[2], Yan Zhou[2] and Guangtao Zhai[1]**

[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China, [2] Department of Radiology, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:
Jun Shi,
Shanghai University, China

Reviewed by:
Mingxia Liu,
The University of North Carolina
at Chapel Hill, United States
Yang Li,
Beihang University, China

***Correspondence:**
Yan Zhou
clare1475@hotmail.com
Guangtao Zhai
zhaiguangtao@sjtu.edu.cn

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: 18 March 2020
Accepted: 06 May 2020
Published: 18 June 2020

Citation:
Chen Q, Wang Y, Qiu Y, Wu X, Zhou Y and Zhai G (2020) A Deep Learning-Based Model for Classification of Different Subtypes of Subcortical Vascular Cognitive Impairment With FLAIR. Front. Neurosci. 14:557. doi: 10.3389/fnins.2020.00557

Deep learning methods have shown their great capability of extracting high-level features from image and have been used for effective medical imaging classification recently. However, training samples of medical images are restricted by the amount of patients as well as medical ethics issues, making it hard to train the neural networks. In this paper, we propose a novel end-to-end three-dimensional (3D) attention-based residual neural network (ResNet) architecture to classify different subtypes of subcortical vascular cognitive impairment (SVCI) with single-shot T2-weighted fluid-attenuated inversion recovery (FLAIR) sequence. Our aim is to develop a convolutional neural network to provide a convenient and effective way to assist doctors in the diagnosis and early treatment of the different subtypes of SVCI. The experiment data in this paper are collected from 242 patients from the Neurology Department of Renji Hospital, including 78 amnestic mild cognitive impairment (a-MCI), 70 nonamnestic MCI (na-MCI), and 94 no cognitive impairment (NCI). The accuracy of our proposed model has reached 98.6% on a training set and 97.3% on a validation set. The test accuracy on an untrained testing set reaches 93.8% with robustness. Our proposed method can provide a convenient and effective way to assist doctors in the diagnosis and early treatment.

Keywords: subcortical ischemic vascular disease, convolutional neural network, deep learning, magnetic resonance imaging, cognitive impairment

INTRODUCTION

Vascular cognitive impairment (VCI) is a broad term that includes a group of cognitive disorders with various degrees of severity, from mild to severe attributable to pathological damage of the cerebral vascular system (Barbay et al., 2017). Vascular dementia developed from VCI is the second most common cause of dementia after Alzheimer's disease (AD) (Barbay et al., 2017). Recently, VCI, especially its most common form subcortical VCI (SVCI), has been getting increased attention, for there is increasing evidence that impaired vascular structure and function are also important in the development of AD (Lucy et al., 2017). SVCI is defined as a clinical continuum of cognitive impairments due to cerebral small vessel disease (Olivia et al., 2018).

Lacunar infarct and white matter hyperintensities (WMHs) (also termed white matter lesions or leukoaraiosis), which are located subcortically or deeply), are the main type of lesions. Prominent perivascular spaces, cerebral microbleeds, and atrophy are the other common signs shown in conventional MRI sequences that are associated with SVCI (Jee and Lee, 2014). Nowadays, with the development of neuroimaging studies, we have gradually found that conventional MRI characteristics cannot fully explain the variable clinical manifestations of SVCI. For example, although voxel-based morphometry and lesion-symptom mapping studies have shown extensive brain damages in SVCI patients, the relationship between these damages and clinical cognitive impairments is still controversial among different studies (Marco et al., 2011; Biesbroek et al., 2013, 2017). The International Society for Vascular Behavior and Cognitive Disorders suggested that a strategic infarct or hemorrhage, multiple lacunes, one large infarct or hemorrhage, and extensive and confluent WMH of vascular origin may be helpful in the diagnosis of SVCI (Perminder et al., 2014). However, as there is little validation of these thresholds, the exact clinical relevance patterns for individual patients remain to be discussed. So by now, the diagnosis of SVCI still relies on scrupulous clinical assessment such as detailed medical history enquiry, physical and neurobiological exams, and neuropsychological evaluation, which are costly, are time-consuming, are subjectively dependent, and may even be traumatic. More effective methods to classify and evaluate the cognitive impairments of SVCI are needed.

Noticeably, a small number of studies have made an effort to resolve the dilemma by traditional machine learning (ML) based on neuroimaging data. Using hierarchical fully convolutional network (H-FCN), Lian et al. (2020) automatically identified discriminative atrophy local patches and regions in brain structural MRI (sMRI) and achieve state-of-the-art AD versus normal control (NC) and progressive mild cognitive impairment (pMCI) versus stable MCI (sMCI) classification performance. By combining diffusion tensor imaging (DTI) and brain morphometry parameters, Stefano et al. (2015) successfully discriminated healthy controls from patients with vascular dementia and vascular MCI (VaMCI) by ML techniques. Stefano et al. (2016) adopted a support vector machine (SVM)-based ML strategy for discrimination SVCI patients with different cognitive performances on the basis of predefined feature vectors extracted from DTI data. The sensitivity, specificity, and accuracy of the classification model were 72.7–89.5%, 71.4–83.3%, and 77.5–80.0%, respectively. Finally, except for not being sensitive enough, extracting those features on the basis of such large data volume of neuroimaging needs human experts, which are often costly, time-consuming, and burdensome. Deep learning (DL) is a rapidly developing ML algorithm for directly extracting high-throughput features from the images without the engagement of human experts. In particular, quite a lot of studies focus on the application of DL-based diagnosis assistance system. Duan et al. (2019) have researched the visual attention analysis of children with autism spectrum disorder (ASD). Liu et al. (2019) focused on the AD diagnosis and used deep multi-task multi-channel learning to achieve state-of-the-art classification results. Yang et al. (2020b) fused deep spatial and temporal features from adaptive dynamic functional connectivity (dFC) and achieved great classification accuracy of 87.7%, which is 5.5% higher than that of the state-of-the-art methods. Liu et al. (2018) proposed a deep multi-instance convolutional neural network (CNN) to automatically learn both local and global representations for MR images and achieve superior performance over state-of-the-art approaches. In particular, in MCI classification problems, Yang et al. (2014, 2020a) have proposed effective sparse functional connectivity networks and sparse multivariate autoregressive modeling methods for MCI classification. In our previous study (Yao et al., 2019), we trained a CNN to classify different cognitive performances in patients with subcortical ischemic vascular disease (SIVD) on the basis of T2-weighted fluid-attenuated inversion recovery (FLAIR) data. For the three-dimensional (3D)-based model, the accuracy of a training set and a testing set reached 99.7 and 96.9%, respectively. This previous study suggests us that DL, especially 3D-CNN, is a powerful and convenient method for classification of SVCI by single-shot T2-weighted FLAIR sequence. By focusing on the sparse regression of blood oxygenation level dependent (BOLD) MRI and arterial spin labeling (ASL) MRI as well as the brain connectivity network inferred from the MR image, Li et al. (2019) and Yang et al. (2019) proposed novel state-of-the-art methods on MCI classification.

With the successful use of 3D-CNN in classifying different stages of cognitive impairment in SVCI, we decided to further our study and refine the model for classifying different subtypes of VaMCI on the basis of the single-shot FLAIR sequence. VaMCI is an intermediate and reversible state between normal cognitive status and vascular dementia. The definition of MCI according to criteria proposed by a multidisciplinary and international experts group includes four clinical subtypes: amnestic MCI (a-MCI; single or multiple domain) and nonamnestic MCI (na-MCI; single or multiple domain) (Winblad et al., 2004). Different VaMCI subtypes might subtend different etiologies: a-MCI (single or multiple domains) was considered to have a degenerative etiology, and multidomain MCI (either amnestic or not) was considered to have a vascular etiology (Emilia et al., 2016). The subtypes of VaMCI are important for clinical care and targeted treatment and might be associated with prognosis. David et al. (2015) found that dementia risks were higher for a MCI than for na-MCI, and for multidomain compared with single-domain MCI.Liesbeth et al. (2017) found that the relevance of reversion for progression risk depends on the MCI subtype. The risk of dementia in participants with MCI who did not revert, especially in amnestic subtype, was higher than in reverters. Neuroimaging studies showed some signs in differentiating a-MCI and na-MCI. Yukako et al. (2019) found that medial temporal lobe atrophy and lower educational history are quick indicators of amnestic cognitive impairment after stroke. Another study showed that medial temporal lobe atrophy was more frequent in multidomain compared with single domain (Emilia et al., 2016). Hosseini et al. (2017) compared different subtypes of VCI on the basis of DTI and FLAIR data. Results showed that higher medial temporal lobe atrophy and left hippocampal mean diffusivity contributed to amnestic VCI and that higher ischemic burden contributed to nonamnestic VCI.

Considering the importance of VaMCI subtypes for clinical decision, and the possibility for image classification suggested by limited neuroimaging studies, we constructed an efficient 3D-CNN model to achieve accurate classification of VaMCI subtypes. To our knowledge, no similar studies have been reported.

MATERIALS AND METHODS

Participants

A total of 242 subjects with SIVD were recruited from patients admitted to the Neurology Department of Renji Hospital from July 2012 to January 2018. SIVD is defined as subcortical WMH on T2-weighted images with at least one lacunar infarct, in accordance with the criteria suggested by Galluzzi (Samantha et al., 2005). All participants received baseline evaluation, including complete collection of sociodemographic and clinical (cognitive, behavioral, neurological, functional, and physical) data. Patient histories were collected from knowledgeable informants, usually from their spouses. All patients underwent laboratory examinations and conventional MRI for routine investigation (Yao et al., 2019).

The exclusion criteria (Yao et al., 2019) were cerebral hemorrhages, cortical and/or corticosubcortical non-lacunar territorial infarcts and watershed infarcts, specific causes of white matter lesions (e.g., multiple sclerosis, sarcoidosis, and brain irradiation), neurodegenerative disease (including AD and Parkinson's disease), and signs of normal pressure hydrocephalus or alcoholic encephalopathy. Patients with low education level (<6 years), severe depression [Hamilton Depression Rating Scale (HDRS) $\geq$ 18], other psychiatric comorbidities or severe cognitive impairment (inability to perform neuropsychological tests), severe claustrophobia, and contraindications to MRI (e.g., pacemaker and metallic foreign bodies) were also excluded. All the participants had lacunar infarcts, small white matter hyperintensities, and slight atrophy.

Finally, all SIVD patients recruited were subdivided based on cognitive status into subcortical vascular disease with no cognitive impairment (NCI) group ($n = 94$) and VaMCI group ($n = 148$). All the participants were right-handed.

The current study was approved by the Research Ethics Committee of Renji Hospital, School of Medicine, Shanghai Jiao Tong University, China. Written informed consent was obtained from each patient.

Neuropsychological Assessment

Neuropsychological assessments (Yao et al., 2019) were performed within 2 weeks of the MRI. All subjects did not suffer a new clinical stroke or TIA between the MRI and assessment. A comprehensive battery of neuropsychological tests was designed based on a review of relevant published reports. These tests are as follows: Trail-Making Tests A and B, Stroop color–word test, verbal fluency (category) test, auditory verbal learning test (short and long delayed free recall), Rey–Osterrieth Complex Figure Test (delayed recall), Boston Naming Test (30 words), Rey–Osterrieth Complex Figure Test (copy), Lawton and Brody's Activities of Daily Living (ADL) Scale Test, Barthel index (BI), HDRS, and the Neuropsychiatric Inventory.

To assess the cognitive status of subjects, the scores for each measure of normal-aged patients in Shanghai, China, were used as the normal baseline (norms) (Yao et al., 2019). Cognitive dysfunction was defined as −1.5 SD in at least one neuropsychological test. According to the AHA Statement on Vascular Contributions to Cognitive Impairment and Dementia (Philip et al., 2011), VaD diagnosis was based on a decline in cognitive function from a prior baseline and a deficit in performance in ≥2 cognitive domains that were of sufficient severity to affect the subject's activities of daily living, which were independent of the motor/sensory sequelae of the vascular event. VaMCI diagnosis was based on the following criteria: (1) ADL could be normal or mildly impaired, (2) does not meet criteria for dementia, and (3) mild quantifiable cognitive impairment within one or more domains (i.e., attention, executive function, memory, language, and visuospatial function). Functional ability was assessed using BI and Lawton and Brody's ADL scales. However, because most patients with cognitive impairment due to cerebrovascular disease have some degree of disability, the study carefully excluded those with disability due to cognitive damage and motor sequelae using cognitive impairment history and clinical judgment. The definition of subtypes of MCI according to criteria proposed by a multidisciplinary and international experts group includes a-MCI and na-MCI (Winblad et al., 2004). NCI was defined as subcortical vascular disease with NCI, which means their scores in all neuropsychological tests were within the normal range (<-1.5 SD).

MRI Protocol

MRI was performed with the SignaHDxt 3T MRI scanner (GE Healthcare, United States). An eight-channel standard head coil with foam padding was used to restrict head motion. Besides conventional brain MRI plain scanning, T2-weighted FLAIR sequences with high resolution were acquired as follows: TE = 150 ms, TR = 9,075 ms, TI = 2,250 ms, field of view (FOV) = 256 × 256 mm^2, matrix = 128 × 128, slice thickness = 2 mm, number of slices = 66.

MRI Data Preprocessing Pipeline

In this section, we propose an end-to-end data pipeline for MR image data processing. The data pipeline contains data preprocessing and model training. Our raw data are T2-weighted FLAIR MR image collected from 242 patients including 78 a-MCI, 70 na-MCI, and 94 NCI. We split the total dataset to three parts including a training set, a validation set, and a testing set with percentage of 60, 20, and 20%, respectively. **Figure 1** shows our proposed MRI data processing pipeline. First, we process the raw data using our data preprocessing method and get trainable data as the input of CNN. Then we feed these processed data into our proposed 3D deep residual network to extract higher-level features and carry out the classification procedure. In the following two sections, we will introduce the pipeline in detail. processing pipeline.

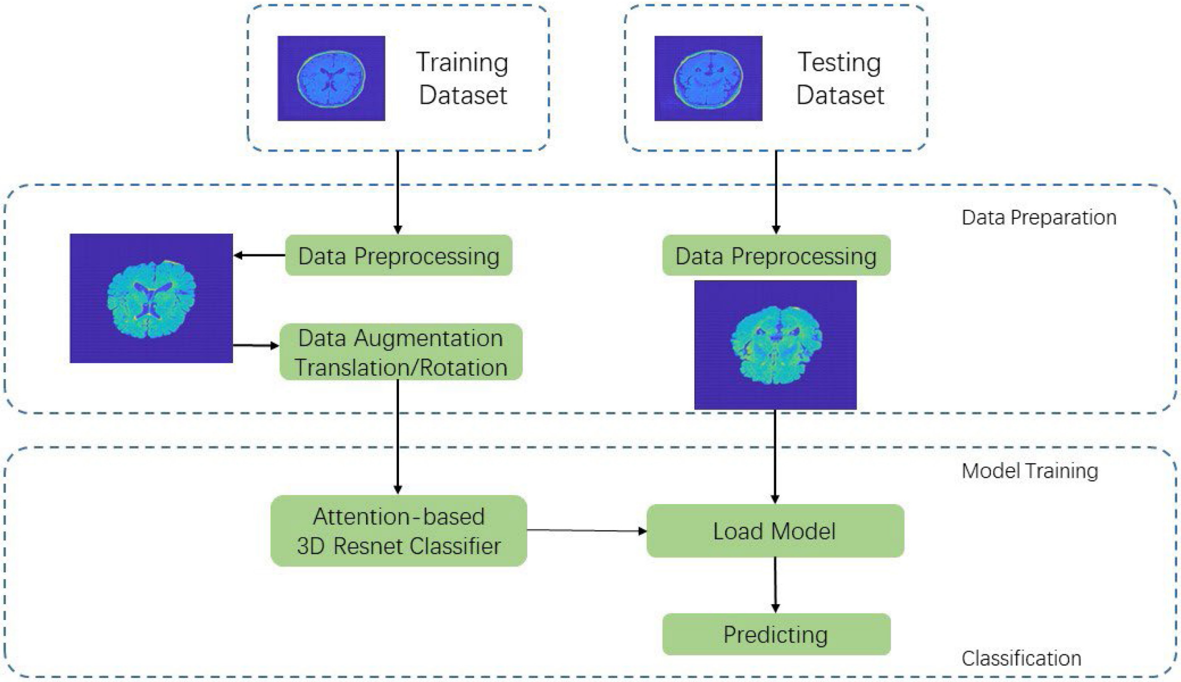


FIGURE 1 | MRI data processing pipeline.

Data Preprocessing

Space Conversion

The MRI data are acquired by tomography. It always takes a long time to complete the acquisition of MR images from a patient, and the patient will inevitably move during such a long acquisition procedure. These collected raw tomographic MRI data may not be mapped one by one when aligned and cannot be connected between different slices for effective analysis. Thus, we first process the MRI data into the same data coordination to map different slice layers into standard space. In this paper, we use SPM software and MRIcro software in Matlab toolkit to process these raw MR image data. The specific steps include format conversion, slice timing, head movement realignment, image matching, brain segmentation, spatial standardization, and so forth.

Brain Separation Using FSL-BET

Traditional sMRI data contain the total brain scanning data including the skull and other non-brain parts, which is meaningless for convolutional networks to extract features. In this case, the skull and non-brain parts act as random noise, and we need to separate them from brain data. In the specific preprocessing process, we used FSL-BET tool to extract the brain structure. We set the fractional intensity threshold to 0.3 and the vertical gradient in fractional intensity threshold to 0.2. The skull separation processing result diagram is shown in **Figure 2**.

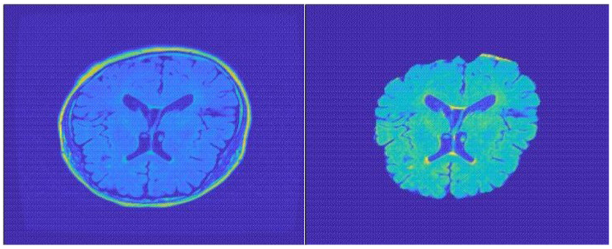


FIGURE 2 | Top view of MR image: the **left** one is before separation; the **right** one after separation.

Brain Region of Interest Segmentation

We transform the DICOM FLAIR image into mat format in MATLAB with the shape of $l \times w \times d \times c$ equaling to $256 \times 256 \times 66 \times 1$, where l, w, d, and c represent the length, width, depth, and color channels of the image, respectively. Considering that there are still lots of meaningless zeros surrounding the brain region, we define the nonzero brain region as our region of interest (ROI) and use contour finding algorithm to find the maximum ROI part in all slices of samples. We then cut the brain ROI into the size of $159 \times 141 \times 66$. By cutting the ROI, we can focus more on the useful brain region. We can also effectively reduce the number of convolution network parameters, which can speed up the training process as well as reduce the risk of overfitting.

Image Smoothing

Noise cannot be completely avoided under any circumstances, and it is similar for medical images. The main noise sources of MR images are thermal/electrical noise and random noise.

The most common preprocessing method is to filter the image. In this paper, we use smoothing method in SPM software. We use Gaussian filtered convolution kernel function to convolve the spatial domain of the MR image, so as to remove the high-frequency noise part of the image, leaving the corresponding low-frequency blood oxygen level and other signals in the MR image. Through image smoothing, differential errors in the signal caused by the image capacity and structure of different subjects are eliminated.

Data Augmentation

Because the collection of MR images is cumbersome and involves medical ethics issues, the total number of samples in our experiment is 242 and need to be separated into train, validation, and test datasets during model training. The features learned by the model may not have extensiveness and may have serious overfitting problem. In order to solve such problems, this paper refers to the method of data augmentation, which is commonly used for natural images, and it adopts specific-augmentation method for the T2-weighted FLAIR MRI data in this paper. In the data preprocessing process of DL, traditional data augmentation methods mainly aim at the samples of two-dimensional (2D) natural images may have some jitter, noise, and other deviations during the acquisition process. In order to standardize the image, they perform geometric transformation such as translation, flip, rotation, and other augmented transformation. As mentioned above, patients' head may have slightly shift or rotation in the data acquisition process. Thus, in our experiment, image panning and slight rotation augmentation method are used to augment samples in the training set.

Convolutional Neural Networks

Medical images are different from traditional natural images in terms of data dimensions and data representation. With the continuous improvement of medical image collection methods and data storage capabilities, the complexity of medical images at the professional level is also increasing. Previously, medical images could only be used as an auxiliary tool for the subjective decision of doctors. Under current situation of increasing density of medical image data, doctors' experience and ability to judge medical images are difficult to keep up with the pace of image development. However, diagnosis is still based on a traditional knowledge system nowadays.

These advances in medical image data collection have not been applied to clinical diagnosis well, and there is redundancy in medical resources. Thus, it is in great demand to develop new automated clinical diagnosis methods. Previously, a solution to this phenomenon was to use ML to perform prediction, segmentation, diagnosis, and so forth, to realize automated diagnosis process. However, the learning capabilities and models of traditional ML methods are often insufficient to handle such a large number of medical images and high-dimensional data. With the improvement of DL (Lecun et al., 2015) and CNN (Lawrence et al., 1997) and the continuous innovation of computer computing capabilities, a combination of high-performance computers and DL methods can be used to learn and process large-scale medical image data extracted from medical image data and inherent higher-order features of the images.

Network Structure

In natural image processing, CNN generally use 2D kernels to implement feature extraction because natural images are mostly 2D. However, MR images are continuous between different slices from the top to bottom. Given that we do not know the exact lesion area of SVCI disease, we use combination of 3 × 3 × 3 and 7 × 7 × 7 three-dimensional convolutional kernels instead of using traditional 2D convolutional kernels to extract 3D features.

Our network uses residual neural network (ResNet)-18 (He et al., 2016) as backbone, which has the best classification effect in 2D natural images and change the structure of the convolution kernel in the model into 3D convolutional kernels so that it can be used for the classification of 3D MR images.

Considering the high density of MRI data in this experiment, our network has a larger number of parameters and a smaller sample size to train this model, which makes it difficult for convergence during the training process. We are inspired by the attention mechanism (Vaswani et al., 2017; Jin et al., 2019) and propose an end-to-end attention-based 3D ResNet model for classification of different subtypes of SVCI on the basis of T2-weighted FLAIR MR images.

Attention model in DL simulates the human brain. When a person is observing a picture, although his or her receptive field can see the entire area of the image, his or her attention to the entire image is not balanced. There is a certain weight to distinguish different regions in human vision, and the effective area that the eyes focus on is actually a very small part. In our experiment, high-density MR image will produce more parameters in neural network. If a model wants to memorize more information of the input image, it has to increase the complexity of the network, which will produce more parameters. This will be a huge burden to our compute capability. Thus, in this paper, we import attention module into our network to focus more on the important region to classify different subtypes of SVCI. In this paper, we use a 3 × 3 × 3 convolution filters activated by ReLU as a subway after convolution feature maps $F_{i,c}$ to produce our attention mask A_i. We then multiply attention mask A_i to previous feature maps $F_{i,c}$, so that we can get the weighted attention map M_{ic} by the following equation:

$$M_{ic} = A_i * F_{ic}$$

The attention mask A_i can be trained and optimized through model training to focus more on the significant parts. Our proposed network structure is shown in **Figure 3**. The network is composed of convolutional layers, ResNet blocks, attention blocks, and output classifier. For example, the Conv3D thirty-two 3 × 3 × 3 strides = 1 layer means 32 convolution filters with the size of 3 × 3 × 3 and strides equal to 1. Different from 2D convolution filters, these filters can receive data from three adjacent slices and can extract features between slices. We fed our preprocessed data with resolution of 159 × 141 × 66 × 1 into the network and go through eight residual blocks. As the layers go deeper, the numbers of filters will increase from 32 to 256, and the features extracted will be more abstract and complex.

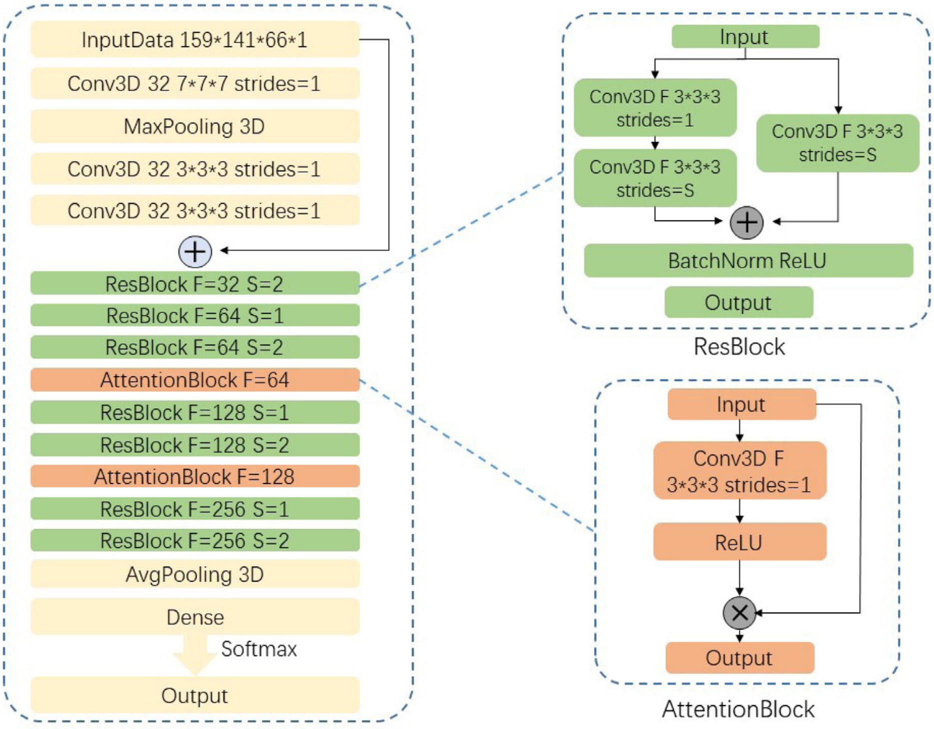


FIGURE 3 | Network structure.

Correspondingly, the last layers in ResBlock have parameter S set to 2, which means that we set the strides to 2 × 2 × 2 and downsample the feature maps size by two times. Then the last output feature maps will be average-pooled and fed into the classifier.

Experiment Settings

We implement the experiment on two NVIDIA GTX 1080 Ti GPUs. We applied the k-fold cross-validation method in training. The total dataset are divided into five equal shares; and for each training process, we use four shares as training and validation sets and one share as the testing set. The final test accuracy and other metrics are calculated by the average of five experiments. The experiment is based on Keras using TensorFlow as backend. Limited by the computation ability, our batch size is set to 4. In our network, preprocessed data with the shape of 159 × 141 × 66 × 1 are fed and are filtered by gradually increasing filters to extract high-level features. The features are finally fed into a fully connected (FC) layer activated by softmax to get the final classification output. We use cross-entropy loss function and adaptive gradient algorithm (Adagrad) optimizer to help our model minimize the loss function. Cross-entropy loss function is shown as follows:

$$L_{\log}(Y,P) = -\log \Pr(Y|P)$$

$$= -\frac{1}{N}\sum_{i=1}^{N-1}\sum_{k=0}^{K-1}(y_{ik})\log(p_{ik})$$

where multivariate classification **k** is the total number of categories, y_{ik} equals to 1 only if the label of the i-th sample is in category **k**, the true category label of **N** samples is an $N \times k$ matrix **Y**, and the probability of each sample in **N** samples predicted by the classifier is an $N \times k$ matrix **P**.

The updated formula of Adagrad is shown below:

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii}+\varepsilon}}$$

where g is the gratitude at time θ_i; in our experiment, we set η as 0.01. Adagrad can do larger updates for low-frequency parameters and smaller updates for high-frequency and can solve the problem that different parameters cannot be updated to different scales according to the importance of the parameters.

RESULTS

In our experiment, we train the proposed attention-based 3D ResNet for 50 epochs. Because there are no relative pretrained models in our classification of different subtypes of SVCI with FLAIR MR image, we train our model with random initialization. With proper hyper-parameter tuning, we approach the best performance on the training set and validation set as shown in **Figure 4**.

Because there are no such methods for the classification of different subtypes of SVCI, our proposed model has significant clinical value. The accuracy of our proposed model on the testing

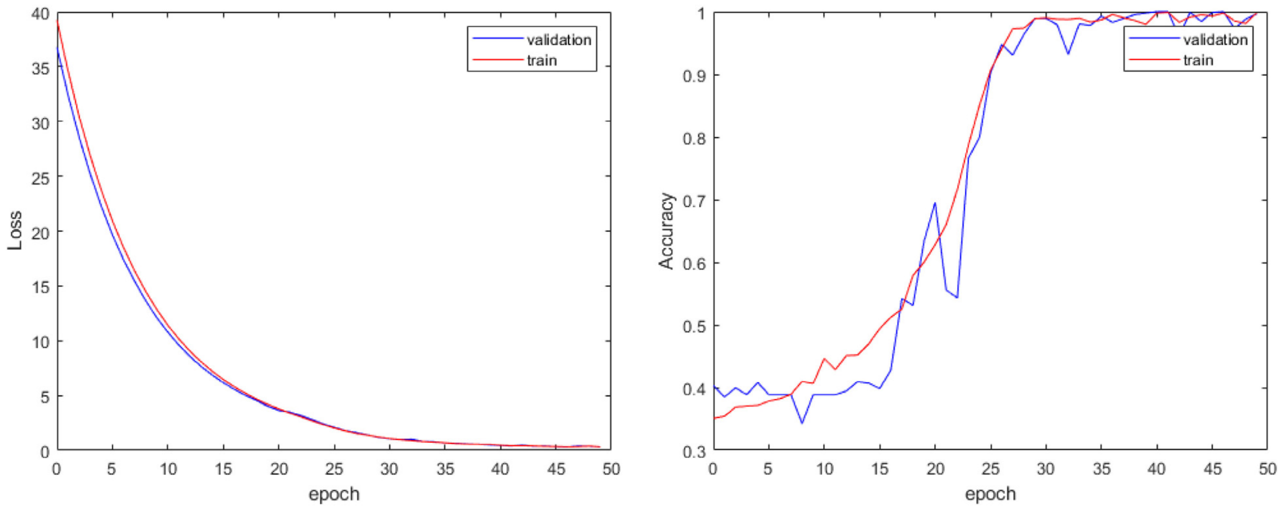


FIGURE 4 | Loss curve and accuracy curve of our model.

set reaches 93.8% with robustness. Thus, our proposed method can effectively assist doctors in early detailed classification diagnosis, so as to carry out targeted treatment in time. In addition to test accuracy, we also consider three other indexes for comparison. We introduce recall (R), precision (P), and F1 score. Precision is the ratio of true positive samples to predicted positive samples, and recall is ratio of true positive samples to actual positive samples. These two indexes can be combined to F1 score as a thorough evaluation of the classification. The formulas of these three indexes are as shown below:

$$R = \frac{TP}{TP + FP}$$

$$P = \frac{TP}{TP + FN}$$

$$\frac{2}{F1} = \frac{1}{R} + \frac{1}{P}$$

where TP, FP, FN, and TN represent positive samples classified to positive, negative samples classified to positive, positive samples classified to negative, and negative samples classified to negative, respectively. Because our experiment is a three-category classification problem, we consider one category as positive samples and the other two as negative samples each time.

The final performance of the model is shown in **Table 1**:

TABLE 1 | Three other index performance of proposed method under three subtypes of subcortical vascular cognitive impairment.

Subtypes	Recall/%	Precision/%	F1 score/%
A-MCI	93.2	91.9	92.6
NA-MCI	94.3	94.3	94.3
NCI	93.8	94.7	94.2

DISCUSSION

Using 3D convolutional kernels, we successfully trained an efficient CNN model that could accurately classify different subtypes of VaMCI (a-MCI and na-MCI) as well as NCI by extracting 3D features from raw T2-weighted FLAIR brain scans. The accuracy of the training set and the testing set reached 98.9 and 97.3% after 50 epochs, respectively. It furthered our previous work of classifying different cognitive performances in SIVD, which is also based on single FLAIR sequence (Yao et al., 2019). These two studies together proved that the method of 3D CNN combined with high-resolution sMRI was worth applying in clinical evaluation of small vessel disease in the elderly. FLAIR sequence was used in our study because it could maximally reflect the imaging features of SVCI such as lacunar infarct and WMH, and the result finally verified the validity of the sequence.

Nowadays, neuroimaging examination has become an indispensable part of clinical evaluation in SVCI, especially MRI with a variety of advanced sequences such as DTI, susceptibility-weighted imaging (SWI), functional MRI, and perfusion-weighted imaging. However, as a result of the imbalance of patients' benefits from the expensive and time-consuming MRI examination, there is still a lack of methods worthy of promotion for the accurate diagnosis and evaluation of patients. DL offered us an opportunity to obtain high clinical diagnostic accuracy with even one single sequence, for it can take full advantage of spatial contextual information in MRI volumes to extract more representative high-level feathers. It could greatly shorten the MRI examination time, reduce the patient's stress caused by the long-time examination, avoid the use of a large number of expensive advanced MRI sequences, and simplify the complex and time-consuming postprocessing. It is important to note that in order to get high-quality image information, we collected high-resolution FLAIR images, which cost 6 min 30 s. Whether thick-layer images as a clinical diagnosis most often used could achieve similar accuracy needs further research.

Considering that high-resolution MR imaging data consist of numerous slices that have a continuous spatial positional relationship, we applied a 3D-based CNN model rather than a 2D-based network, which has been proved to be more efficient in our previous study (Yao et al., 2019). Finally, we got a high accuracy of subclassifying VaMCI into a-MCI and na-MCI.

The subclassification of MCI has clinical significance, because different MCI subtypes may subtend different etiologies. a-MCI may indicate a degenerative etiology and has higher dementia risks than has na-MCI, whereas na-MCI may indicate a vascular etiology that needs more treatment to improve vascular function and cerebral perfusion (Liu et al., 2018, 2019; Yang et al., 2020b). On the basis of single high-resolution FLAIR images, we proved that 3D-CNN can classify not only different cognitive impairment stages in SIVD but also subtypes in MCI stage. This method can greatly improve the efficiency and accuracy of clinical diagnosis of SVCI and is beneficial to clinical targeted treatment at the early stage of cognitive impairment.

Although we have achieved an appealing performance with a high accuracy in this study, there are still several limitations. First, this is a retrospective study with a relatively small sample size. Large-scale multicenter and perspective studies are needed to fully assess the generalization ability of the model. Second, more detailed clinical groups such as single domain and multidomain cognitive groups with or without amnesia based on sufficient sample size can further test this 3D-CNN model and enrich its clinical application. Third, the clinical or pathological interpretation of the association between the high-level features and the cognitive performances remain challenging. Further studies are needed to establish a rationale to explain the correlation between deep imaging features and cognitive performances, which might hint at the underlying pathological mechanisms of SVCI.

CONCLUSION

In this paper, we proposed an end-to-end attention-based 3D ResNet model for classification of different subtypes of SVCI with T2-weighted FLAIR MR images. End to end means doctors do not need to perform complicated data preprocessing; they can simply input the single MRI scanning image of patients to the model and get the output of SVCI classification. Then they can further get the diagnostic decision results according to the auxiliary diagnosis results of our proposed methods. Our proposed method provides a convenient and effective way to assist doctors in the diagnosis and early treatment.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding authors.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Research Ethics Committee of Renji Hospital, School of Medicine, Shanghai Jiao Tong University, China. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YZ and GZ designed and instruct the experiments. QC wrote the code for the experiments. QC and YW carried out the experiments and wrote the manuscript. YW, YQ, and XW collected and analyzed the experiment data. All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

This work was supported by National Natural Science Foundation of China (Award Number(s): 61831015 and U1908210), the Shanghai Municipal Education Commission-Gaofeng Clinical Medicine Grant Support (No. 20172013); Shanghai Science and Technology Committee Medical Guide Project (No. 17411964300); Research Seed Fund of Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University (RJZZ18-003); Incubating Program for Clinical Research and Innovation of Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University (PYIV-17-003 and PY2018-III-03).

REFERENCES

Barbay, M., Taillia, H., Nedelec-Ciceri, C., Arnoux, A., Puy, L., Wiener, E., et al. (2017). Vascular cognitive impairment: advances and trends. *Revue Neurol.* 173, 473–480. doi: 10.1016/j.neurol.2017.06.009

Biesbroek, J. M., Kuijf, H. J., Van Der Graaf, Y., Vincken, K. L., Postma, A., Mali, W. P. T. M., et al. (2013). Association between subcortical vascular lesion location and cognition: a voxel-based and tract-based lesion-symptom mapping study. The SMART-MR study. *PLoS One* 8:e60541. doi: 10.1371/journal.pone.0060541

Biesbroek, J. M., Weaver, N. A., and Biessels, G. J. (2017). Lesion location and cognitive impact of cerebral small vessel disease. *Clin. Sci. (Lond. Engl. 1979)* 131, 715–728. doi: 10.1042/cs20160452

David, S. K., Beiser, A., Machulda, M. M., Fields, J., Roberts, R. O., Pankratz, V. S., et al. (2015). Spectrum of cognition short of dementia: framingham heart study and mayo clinic study of aging. *Neurology* 85, 1712–1721. doi: 10.1212/wnl.0000000000002100

Duan, H., Min, X., Fang, Y., Fan, L., Yang, X., and Zhai, G. (2019). Visual attention analysis and prediction on human faces for children with autism spectrum disorder. *ACM Transact. Multimed. Comput. Commun. Applicat.* 15, 1–23. doi: 10.1145/3337066

Emilia, S., Poggesi, A., Valenti, R., Pracucci, G., Pescini, F., Pasi, M., et al. (2016). Operationalizing mild cognitive impairment criteria in small vessel disease: the VMCI-Tuscany Study. *Alzheimers Dem. J. Alzheimers Assoc.* 12, 407–418. doi: 10.1016/j.jalz.2015.02.010

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

Hosseini, A. A., Meng, D., Simpson, R. J., and Auer, D. P. (2017). Mesiotemporal atrophy and hippocampal diffusivity distinguish amnestic from non-amnestic vascular cognitive impairment. *Eur. J. Neurol.* 24, 902–911. doi: 10.1111/ene.13299

Jee, H. R., and Lee, J-H. (2014). Recent updates on subcortical ischemic vascular dementia. *J. Stroke* 16, 18–26.

Jin, D., Zhao, K., Hu, F., Yang, Z., Liu, B., Jiang, T., et al. (2019). "Attention-based 3D convolutional network for Alzheimer's disease diagnosis and biomarkers exploration," in *Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, 1047–1051.

Lawrence, S., Giles, C. L., Tsoi, A. C., and Back, A. D. (1997). Face recognition: a convolutional neural-network approach. *IEEE Transact. Neural Netw.* 8, 98–113. doi: 10.1109/72.554195

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

Li, Y., Yang, H., Lei, B., Liu, J., and Wee, C. (2019). Novel Effective connectivity inference using ultra-group constrained orthogonal forward regression and elastic multilayer perceptron classifier for MCI identification. *IEEE Transact. Med. Imaging* 38, 1227–1239. doi: 10.1109/tmi.2018.2882189

Lian, C., Liu, M., Zhang, J., and Shen, D. (2020). Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Transact. Patt. Anal. Mach. Intell.* 42, 880–893. doi: 10.1109/tpami.2018.2889096

Liesbeth, A., Heffernan, M., Kochan, N. A., Crawford, J. D., Draper, B., Trollor, J. N., et al. (2017). Effects of MCI subtype and reversion on progression to dementia in a community sample. *Neurology* 88, 2225–2232. doi: 10.1212/wnl.0000000000004015

Liu, M., Zhang, J., Adeli, E., and Shen, D. (2018). 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, M., Zhang, J., Adeli, E., and Shen, D. (2019). Joint Classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Transact. Biomed. Eng.* 66, 1195–1206. doi: 10.1109/tbme.2018.2869989

Lucy, B., Haunton, V. J., Panerai, R. B., and Robinson, T. G. (2017). Cerebral hemodynamics in mild cognitive impairment: a systematic review. *J. Alzheimers Dis.* 59, 369–385. doi: 10.3233/jad-170181

Marco, D., Zieren, N., Hervé, D., Jouvent, E., Reyes, S., Peters, N., et al. (2011). Strategic role of frontal white matter tracts in vascular cognitive impairment: a voxel-based lesion-symptom mapping study in CADASIL. *Brain* 134(Pt 8), 2366–2375. doi: 10.1093/brain/awr169

Olivia, A. S., Black, S. E., Chen, C., Decarli, C., Erkinjuntti, T., Ford, G. A., et al. (2018). Progress toward standardized diagnosis of vascular cognitive impairment: guidelines from the vascular impairment of cognition classification consensus study. *Alzheimers Dem. J. Alzheimers Assoc.* 14, 280–292.

Perminder, S., Kalaria, R., O'brien, J., Skoog, I., Alladi, S., Black, S. E., et al. (2014). Diagnostic criteria for vascular cognitive disorders: a VASCOG statement. *Alzheimer Dis. Assoc. Dis.* 28, 206–218. doi: 10.1097/wad.0000000000000034

Philip, B. G., Scuteri, A., Black, S. E., Decarli, C., Greenberg, S. M., Iadecola, C., et al. (2011). Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the american heart association/american stroke association. *Stroke* 42, 2672–2713. doi: 10.1161/str.0b013e3182299496

Samantha, G., Sheu, C.-F., Zanetti, O., and Frisoni, G. B. (2005). Distinctive clinical features of mild cognitive impairment with subcortical cerebrovascular disease. *Dem. Ger. Cogn. Dis.* 19, 196–203. doi: 10.1159/000083499

Stefano, C., Citi, L., Salvadori, E., Valenti, R., Poggesi, A., Inzitari, D., et al. (2016). Prediction of impaired performance in trail making test in MCI patients with small vessel disease using DTI data. *IEEE J. Biomed. Health Informatics* 20, 1026–1033. doi: 10.1109/jbhi.2016.2537808

Stefano, D., Ciulli, S., Ginestroni, A., Salvadori, E., Poggesi, A., Pantoni, L., et al. (2015). "Multimodal MRI classification in vascular mild cognitive impairment," in *Proceedings of The Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* (Piscataway, NJ: IEEE), 4278–4281.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Adv. Neur. Inform. Proces. Syst.* 5998–6008.

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L. O., et al. (2004). Mild cognitive impairment–beyond controversies, towards a consensus: report of the international working group on mild cognitive impairment. *J. Int. Med.* 256, 240–246.

Yang, L., Liu, J., Gao, X., Jie, B., Kim, M., Yap, P.-T., et al. (2019). Multimodal hyper-connectivity of functional networks using functionally-weighted LASSO for MCI classification. *Med. Image Anal.* 52, 80–96. doi: 10.1016/j.media.2018.11.006

Yang, L., Liu, J., Peng, Z., Sheng, C., Kim, M., Yap, P.-T., et al. (2020a). Fusion of ULS group constrained high- and low-order sparse functional connectivity networks for MCI classification. *Neuroinformatics* 18, 1–24. doi: 10.1007/s12021-019-09418-x

Yang, L., Liu, J., Tang, Z., and Lei, B. (2020b). Deep spatial-temporal feature fusion from adaptive dynamic functional connectivity for MCI identification. *IEEE Transact. Med. Imaging* doi: 10.1109/TMI.2020.2976825

Yang, L., Wee, C.-Y., Jie, B., Peng, Z., and Shen, D. (2014). Sparse multivariate autoregressive modeling for mild cognitive impairment classification. *Neuroinformatics* 12, 455–469. doi: 10.1007/s12021-014-9221-x

Yao, W., Tu, D., Du, J., Han, X., Sun, Y., Xu, Q., et al. (2019). Classification of subcortical vascular cognitive impairment using single MRI sequence and deep learning convolutional neural networks. *Front. Neurosci.* 13:627.

Yukako, T., Saito, S., Yamamoto, Y., Uehara, T., Yokota, C., Sakai, G., et al. (2019). Visually-rated medial temporal lobe atrophy with lower educational history as a quick indicator of amnestic cognitive impairment after stroke. *J. Alzheimers Dis.* 67, 621–629. doi: 10.3233/jad-180976

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 Neuroscience

ORIGINAL RESEARCH
published: 18 September 2020
doi: 10.3389/fnins.2020.00858

Early Prediction of Cognitive Deficit in Very Preterm Infants Using Brain Structural Connectome With Transfer Learning Enhanced Deep Convolutional Neural Networks

Ming Chen[1,2], *Hailong Li*[1], *Jinghua Wang*[3], *Weihong Yuan*[3,4], *Mekbib Altaye*[5,6], *Nehal A. Parikh*[1,6] *and Lili He*[1,6]*

[1] *The Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States,* [2] *Department of Electronic Engineering and Computing Systems, University of Cincinnati, Cincinnati, OH, United States,* [3] *Department of Radiology, University of Cincinnati College of Medicine, Cincinnati, OH, United States,* [4] *Department of Radiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States,* [5] *Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States,* [6] *Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States*

Edited by:
Hyunjin Park,
Sungkyunkwan University,
South Korea

Reviewed by:
Amy Kuceyeski,
Cornell University, United States
Xi Jiang,
University of Electronic Science and Technology of China, China

***Correspondence:**
Lili He
lili.he@cchmc.org

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *17 May 2020*
Accepted: *23 July 2020*
Published: *18 September 2020*

Citation:
Chen M, Li H, Wang J, Yuan W, Altaye M, Parikh NA and He L (2020) Early Prediction of Cognitive Deficit in Very Preterm Infants Using Brain Structural Connectome With Transfer Learning Enhanced Deep Convolutional Neural Networks. Front. Neurosci. 14:858. doi: 10.3389/fnins.2020.00858

Up to 40% of very preterm infants (≤32 weeks' gestational age) were identified with a cognitive deficit at 2 years of age. Yet, accurate clinical diagnosis of cognitive deficit cannot be made until early childhood around 3–5 years of age. Recently, brain structural connectome that was constructed by advanced diffusion tensor imaging (DTI) technique has been playing an important role in understanding human cognitive functions. However, available annotated neuroimaging datasets with clinical and outcome information are usually limited and expensive to enlarge in the very preterm infants' studies. These challenges hinder the development of neonatal prognostic tools for early prediction of cognitive deficit in very preterm infants. In this study, we considered the brain structural connectome as a 2D image and applied established deep convolutional neural networks to learn the spatial and topological information of the brain connectome. Furthermore, the transfer learning technique was utilized to mitigate the issue of insufficient training data. As such, we developed a transfer learning enhanced convolutional neural network (TL-CNN) model for early prediction of cognitive assessment at 2 years of age in very preterm infants using brain structural connectome. A total of 110 very preterm infants were enrolled in this work. Brain structural connectome was constructed using DTI images scanned at term-equivalent age. Bayley III cognitive assessments were conducted at 2 years of corrected age. We applied the proposed model to both cognitive deficit classification and continuous cognitive score prediction tasks. The results demonstrated that TL-CNN achieved improved performance compared to multiple peer models. Finally, we identified the brain regions most discriminative to the cognitive deficit. The results suggest that deep learning models may facilitate early prediction of later neurodevelopmental outcomes in very preterm infants at term-equivalent age.

Keywords: convolutional neural network, deep learning, cognitive deficit, transfer learning, structural connectome

INTRODUCTION

A high prevalence of long-term cognitive deficit is well-established in very preterm infants (≤32 weeks' gestational age), with 35–40% of this population identified with a deficit at 2 years of age (Blencowe et al., 2012; Hamilton et al., 2016). This neurological deficit may affect the infant throughout life, thereby resulting in difficulties in academic skills and building social relationships. Yet, no robust prognostic screening technique is available following neonatal intensive care stay. Typically, an accurate diagnosis of cognitive deficit cannot be made until early childhood around 3–5 years of age. This delayed diagnosis misses the optimal neuroplasticity period of brain development in the first 3 years of life and potentially undermines the effectiveness of early interventions. As such, reproducible approaches that serve as neonatal prognostic tools are needed to fill the gap in our knowledge about the early prediction of cognitive deficit in very preterm infants.

The human brain is a highly interconnected network with coordinated information transfer among individual brain regions (Sporns et al., 2005). Advanced non-invasive neuroimaging MRI techniques have been applied to construct such network representation of the brain, referred to as the brain connectome (Bullmore and Sporns, 2009). Theoretically, a brain connectome is a graph, where vertices represent a set of brain regions of interest (ROIs) and edges represent brain connectivity between ROIs. This brain connectome perspective shifted traditional research that focuses on isolated ROIs toward research on a systematic mechanism incorporating the whole brain. Brain connectome data have very high dimensionality and are intrinsically complex, creating difficulties in designing feature extraction methods and building analysis models. Deep learning has shown great promise in deciphering complex and high dimensional data (e.g., images, signals, and videos) to achieve superior performance in numerous fields, including computer vision, speech recognition, and natural language processing (LeCun et al., 2015). Indeed, numerous studies have applied deep learning approaches to brain connectome for various neurological disorders (Wee et al., 2012; Barkhof et al., 2014; He et al., 2018; Heinsfeld et al., 2018; Li et al., 2018; Sen et al., 2018; Chen et al., 2019).

Brain connectome plays an important role in understanding human cognitive functions (Nagy et al., 2004; Park and Friston, 2013; Petersen and Sporns, 2015). Recent research demonstrated that deep learning models were capable of predicting later cognitive deficits for neonates using brain structural connectome that was constructed by diffusion tensor imaging (DTI) data (Kawahara et al., 2017; Girault et al., 2019). One method to apply deep learning models to brain connectome data is to ignore the topology of the connectome and reshape the adjacency matrix into a vector of features as input (Munsell et al., 2015; Girault et al., 2019). However, the spatial locality (i.e., 2D grid regions of an adjacency matrix) and topological locality information (i.e., rows/columns of an adjacency matrix) in the brain connectome are not utilized, thereby resulting in information loss and potentially compromising the performance of prediction models. Another approach is to apply specialized topological row and column filters on the adjacency matrix of the brain structural connectome to learn the topological relationship between edges (Kawahara et al., 2017). This approach, however, only emphasizes topological locality and discards the spatial locality information (e.g., physically nearby brain ROIs and associated edges) that are intrinsic to any brain ROI parcellation. Since the brain structural connectome is a modular graph that contains clusters of vertices and edges, its adjacency matrix contains hierarchically segregated modules (Park and Friston, 2013). Those topological filters may extract redundant information within connectome modules and may not be efficient for capturing spatial locality. In this work, we consider the adjacency matrix of brain structural connectome as a 2D image and propose to apply established deep convolutional neural networks (CNNs) to learn the spatial and topological information of the brain connectome.

Although deep CNN models have shown promising results on image classification, those models usually require large datasets for model training. In the studies of very preterm infants, available annotated neuroimaging datasets with clinical and outcome information are usually limited and expensive to enlarge, preventing deep CNN to be directly utilized. Transfer learning (TL) may serve as a potential solution to this challenge. Briefly, TL reuses a pre-trained model designed for one task as a starting point for another related task (Bengio, 2012; Samala et al., 2016, 2018; Shin et al., 2016; Azizi et al., 2017; Kooi et al., 2017; Zheng et al., 2018; Bizzego et al., 2019). Raina et al. (2007) proposed a self-taught learning framework that takes unlabeled images to improve the classification performance of their target classification task. Cheng et al. (2019) transferred image features learned from the early stages of Alzheimer's disease (AD) to improve the prediction of AD diagnosis. Gao et al. (2019) reused pre-trained models based on a large-scale natural image dataset and re-trained a deep learning model for classification of brain activity heatmaps derived from task-based functional MRI data. Recently, we applied the TL technique to a deep neural network (DNN) model for cognitive deficit prediction using brain functional connectome data (He et al., 2018). The DNN model was pre-trained using a large number of brain connectome data in an unsupervised fashion and then fine-tuned with brain connectome data from very preterm infants.

In this study, we proposed a TL-enhanced deep CNN (TL-CNN) model for early prediction of cognitive deficit at 2 years of age in very preterm infants using brain structural connectome derived from at term DTI data. Specifically, the proposed model contains two modules, a very deep CNN (which was trained with supervision using ∼1.2 million images from the ImageNet database) (Deng et al., 2009) and a "shallow" CNN. With the fixed weighted pre-trained very deep CNN, we only need to train and fine-tune the "shallow" CNN using available very preterm infants' brain connectome data and associated risks of cognitive deficit. For individual very preterm infants, we constructed brain structural connectome using mean fractional anisotropy from their DTI data collected at term-equivalent age. The proposed model is able to evaluate at term whether or not a very preterm infant will have a high risk to develop later cognitive deficits as well as to predict this infant's cognitive assessment [standardized

Bayley Scales of Infant and Toddler Development, Third Edition (Bayley III) cognitive score] at 2 years of age.

MATERIALS AND METHODS

Subjects

The study includes a cohort of 110 very preterm infants, born at 31 weeks gestational age or less from four academic and non-academic centers in Columbus, Ohio, including Nationwide Children's Hospital (NCH), Ohio State University Medical Center, Riverside Hospital, and Mount Carmel St. Ann's Hospital. Infants were enrolled between December 2014 and April 2016. All subjects with any congenital or chromosomal anomalies affecting the central nervous system were excluded. Infants with cyanotic congenital heart disease were also excluded. The study was approved by the Institutional Review Board of NCH. Approval at the other hospitals was obtained through reciprocity agreements that were in place with NCH. Written informed consent was obtained from parents or legal guardians of all infants.

MRI and Cognitive Outcome Acquisition

Very preterm infants in the cohort were scanned on a 3T scanner (Skyra; Siemens Healthcare) at NCH using a 32-channel phased-array head coil. The imaging was performed after the infant was fed and in natural sleep without sedation. Natus Mini Muffs (Natus Medical Inc., Scan Carlos, CA, United States) and InstaPuffy Silicone Earplugs (E.A.R Inc., Boulder, CO, United States) were employed for MRI noise reduction. DTI was acquired with echo-planar imaging using the following parameters (b800/b2000): repetition time = 6972/5073 ms; echo time = 88 ms; field of view = 160 mm × 160 mm; in-plane resolution = 2 mm × 2 mm; number of slices = 76; slice thickness = 1 mm; 64 non-colinear diffusion-weighted directions; for all images, one volume has no diffusion sensitization; sensitivity encoding factor equates to 2. High-resolution T2-weighted anatomical images were acquired with rapid spin-echo sequence: TR/TE = 7.3/3.4 ms, flip angle = 11°, voxel dimensions 1.0 mm × 1.0 mm × 1.0 mm, scan time = 2:47 min.

All preterm infants received (Bayley-III) test at 2 years corrected age while blinded to DTI data. The Bayley-III cognitive scores are on a scale of 40–160, with a mean of 100 and a standard deviation of 15.

DTI Data Preprocessing

DTI data were preprocessed using FMRIB's Diffusion Toolbox (in the FMRIB Software Library, FSL, Oxford, United Kingdom) following our previously established pipeline (Yuan et al., 2015). Specifically, head motion and eddy current artifacts were mitigated by aligning all diffusion images to the b0 image via an affine transformation. Diffusion tensor reconstruction and brain fiber tracking were performed in the subject's native space using Diffusion Toolkit/TrackVis (Hess et al., 2006; Wang et al., 2007). Diffusion tensor calculation was based on a linear least-square fitting algorithm, and brain fiber tracking was based on a deterministic tracking algorithm (Wang et al., 2007). The fiber tracking uses an angular threshold of 35°. The fiber length threshold was set to 5 mm. The obtained fractional anisotropy maps were harmonized using a batch-effect correction algorithm ComBat (Fortin et al., 2017). We use a neonatal Automated Anatomical Labeling (AAL) brain atlas proposed by Shi et al. (2011). For each subject, the high-resolution T2-weighted images were first registered to the b0 image in the subject's native space and then to the neonatal template space to obtain a transformation matrix. Next, the inverse transformation matrix was used to transform the parcellated ROIs from the template space back to the subject's native space (b0).

Whole-Brain Structural Connectome Construction

A brain connectome is a graph $G = (A, \Omega)$, where vertices Ω represent a set of ROIs, and A is an adjacency matrix of edges that represent brain connectivity between a pair of ROIs. Ninety ROIs were defined based on a neonatal automated labeling atlas (Shi et al., 2011). The weights of structural connectivity between each pair of ROIs were calculated as the mean FA of all voxels along the WM tract constructed between the two ROIs, resulting in a 90 × 90 symmetric adjacency matrix. This was performed using the UCLA Multimodal Connectivity Package (Bassett et al., 2011).

Overview of TL-Enhanced Deep CNN

The proposed model contains two modules, a very deep CNN (which was trained with supervision using ~1.2 million images from the ImageNet database) (Deng et al., 2009) and a "shallow" CNN. In **Figure 1**, we display a two-stage model training procedure in the top two blocks and picture a clinical application in the bottom block, where the proposed model can aid clinicians in the prediction of cognitive deficit using brain structural connectome data. The model training procedure contains two stages: (1) pre-training in the source domain and (2) fine-tuning in the target domain. Specifically, in stage 1, we first pre-trained a deep CNN to learn the basic transferrable image representation (e.g., edges, shapes, etc.) using a large number of color images and associated image labels (source domain). In stage 2, we reused the pre-trained model from stage 1 and fine-tuned the model in the target domain with brain structural connectome and associated cognition deficit outcomes.

Pre-training in the Source Domain

In the source domain, we trained the proposed model to learn transferrable image representation (e.g., edges, shape, and blobs) from diverse objects (e.g., animals, vehicles, human, and natural environments). We defined the task in the source domain as an image classification task. Adjacency matrices of brain connectome are different from those semantic images (dogs, cats, etc.); however, the low-level imaging features (for example straight and curved lines that construct images) are universal to most image analysis tasks. Therefore, the idea behind TL is to treat the pre-trained model as a feature extractor to extract low-level imaging features from the adjacency matrix of a given structural connectome. In this study, we started with

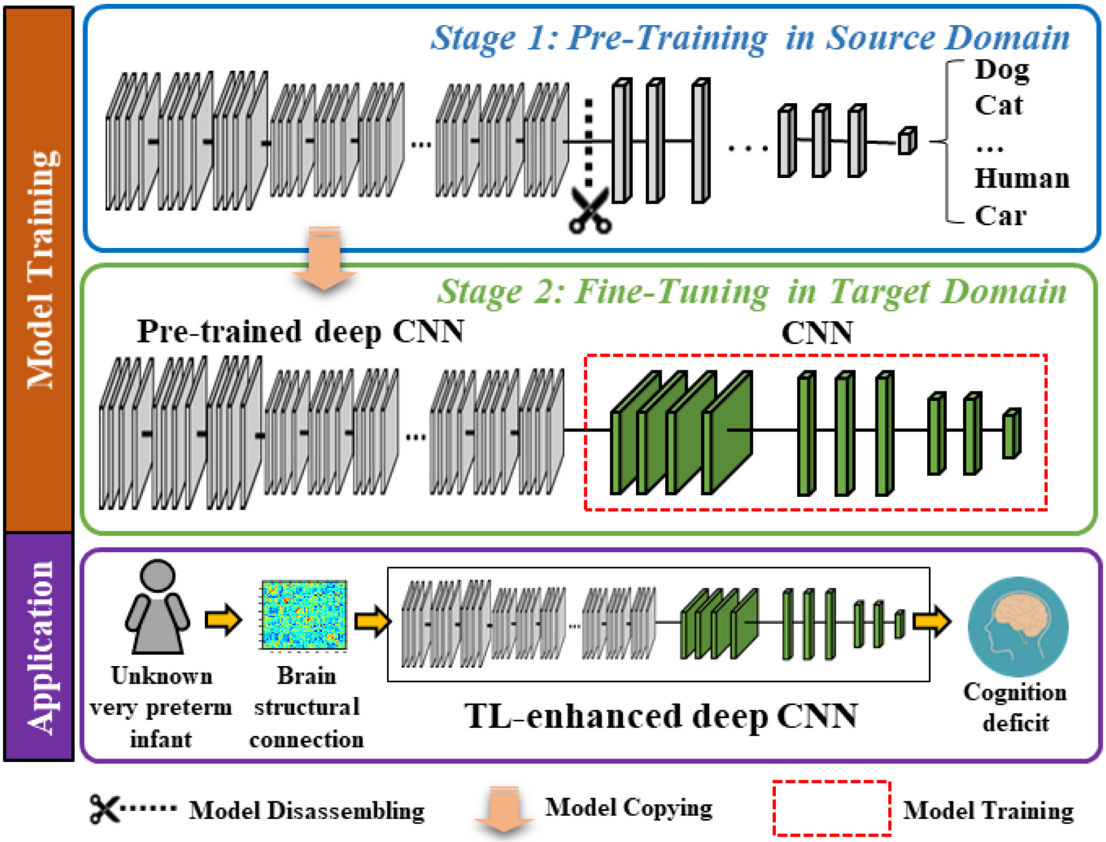


FIGURE 1 | Schematic diagram of the proposed transfer learning-enhanced deep CNN (TL-CNN) model to predict cognitive deficits at 2 years corrected age using brain structural connectome data obtained at term in very preterm infants. The top two blocks demonstrate a two-stage model training procedure, and the bottom block illustrates a potential clinical computer-aided diagnosis application after model training.

the VGG-Nets (Simonyan and Zisserman, 2014) to develop our deep CNN model. VGG-Nets are a set of very deep CNN that were initially proposed by Visual Geometry Group in ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2014. They have been applied to other image analysis applications (Choi et al., 2017; Zhen et al., 2017; Wang et al., 2018). We adopted the architecture of VGG19, one of VGG-Nets models for our study. Briefly, VGG19 is a very deep CNN that consists of 19 trainable layers, including 16 convolutional layers and 3 fully connected (FC) layers designed for classifying 1000 object categories. For each convolutional layer, the VGG19 uses small convolutional filters (3 × 3) along with a rectified linear unit activation function. We obtained a VGG19 model that was pre-trained using ∼1.2 million color images from the ImageNet database. We then dissembled the model and reserved the weights of the convolutional and pooling layers (**Figure 1**, blue box).

Fine-Tuning in the Target Domain

The task in the target domain is to predict the cognitive outcome at 2 years corrected age using brain structural connectome obtained at term-equivalent age. Since the deep CNN in the source domain was pre-trained to recognize transferrable image representation, it would automatically extract image features from the brain structural connectome. The fine-tuning in the target domain is essential to discover discriminative features among generic features and link them to the target task (i.e., cognitive development). We connected a "shallow" CNN (i.e., 2 convolutional layers and 2 FC layers) to the pre-trained fixed weighted deep CNN from the first stage. Finally, an output layer was attached for classification or regression tasks (**Figure 1**, green box). We used brain structural connectome and follow-up cognitive outcomes to fine-tune the deep CNN model. Given N training samples $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_{N-1}, \boldsymbol{x}_N)$ from the target cohort as well as their labels $(y_1, y_2, \ldots, y_i, \ldots, y_{N-1}, y_N)$, where $\boldsymbol{x}_i$ is the i-th input sample (i.e., brain structural connectome) and y_i is the corresponding label, we defined the cross-entropy loss function as:

$$J(\mathrm{W}, \boldsymbol{b}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(p(\boldsymbol{x}_i)\right) + \left(1 - y_i\right)\log\left(1 - p(\boldsymbol{x}_i)\right)$$

where $p(\boldsymbol{x}_i)$ is the predicted probability of $\boldsymbol{x}_i$, W is the weight matrix, and $\boldsymbol{b}$ denotes the bias of the model. In addition to the dichotomized prediction (i.e., classification), we also trained our model to perform continuous cognitive score prediction (i.e., regression). We applied a linear unit at the end of the model and optimized a weighted mean absolute error (MAE) loss function as follows:

$$L(\boldsymbol{W}, \boldsymbol{b}) = \frac{1}{N} \sum_{i=1}^{N} \left| \left(y_i - \hat{y}_i \left(\mathrm{W}, \boldsymbol{b} \right) \right) \right|$$

where $\hat{y}_i (W, \boldsymbol{b})$ is the output of the linear unit of the model, i.e., the predicted score. Similar to the previous cross-entropy loss function, $\boldsymbol{b}$ represents the bias, and W is the weight of the model. The proposed model was optimized using Adam (Kingma and Ba, 2014), a backpropagation gradient descent algorithm. Adam computes adaptive learning rates for weight updating based on the average of recent magnitudes of the gradients, improving computational efficiency. The initial learning rate is set to 0.001. We applied 50 epochs to train the TL-CNN model. The detailed architecture of the TL-CNN model is elaborated in **Supplementary Figure 1**.

Alternative Model Comparison

Linear/Logistic Regression Model

In the linear regression model, we applied mean squared error as the loss function to minimize the residual sum of error between the true score and the score predicted by the linear approximation. For the logistic regression (LR) model, we adopted cross-entropy as the cost function. We used L2 regularization as the penalty term, and we grid searched the regularization parameters with empirical values ($10^{-3}, 10^{-2}, \ldots, 10^{1}$).

Support Vector Machine

We tested the support vector machine (SVM) model with three different kernels: linear, polynomial, and radial basis function (RBF), where the SVM with linear kernel achieved the best prediction performance. Specifically, for all SVM models, we used L2 regularization as the penalty. We grid searched the regularization parameters with empirical values (10^{-3}, 10^{-2},. . ., 10^{1}) and the soft margin parameter C with empirical values (2^{-3}, 2^{-2},. . ., 2^{3}) to optimize the prediction performance. For the polynomial and RBF SVM model, we set the scale gamma kernel coefficient as 1.

Deep Neural Network

The DNN model has an input layer, two FC layers with 256 and 64 neurons, and an output layer. The rectifier linear unit as activation function was used in each neuron. We attached a batch normalization layer and a dropout layer after each FC layer. In the output layer, we used a SoftMax classifier for classification and a linear classifier for regression. The DNN was trained in a supervised fashion and tested using the labeled subjects from the target domain.

TL-DNN

The TL-DNN model has the same structure as the DNN model. Instead of training from scratch, we pre-trained the TL-DNN model in an unsupervised fashion using 257 full-termed neonatal subjects from the source domain. Then, we fine-tuned the TL-DNN with supervision using the labeled subjects from the target domain.

Convolutional Neural Network

The CNN model has two convolutional layers, where each has 256 neurons with a 3 × 3 convolutional filter, and two FC layers, where each layer contains 256 and 64 neurons. A rectified linear unit was used as an activation function. A batch normalization and a dropout layer are attached after each FC layer. We applied a SoftMax classifier for the classification task and linear function for the regression task. The architecture design of this model represents a standard "shallow" CNN model without TL strategy. The CNN model was trained and tested using the subjects from the target domain.

Data Augmentation

The number of very preterm infants in the study cohort is relatively small and imbalanced (i.e., only a small portion of the cohort are at high risk for cognitive deficit). We utilized the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) to balance and augment the training set. Specifically, the training subjects were divided into five bins according to their scores (<70, 70–80, 80–90, 90–100, and >100). Given a bin, a sample was randomly chosen. Then, k nearest neighbors for the selected sample were searched. We set $k = 5$ in this work. A synthetic sample x_{syn} is calculated using the randomly selected sample and its associated neighbors $x_1, x_2, x_3, x_4, x_5, x_6$ by: $x_{syn} = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x_6$, where w_1, w_2, w_3, w_4, w_5, and w_6 are random numbers and $w_1 + w_2 + w_3 + w_4 + w_5 + w_6 = 1$. Similarly, the label y_{syn} for x_{syn} was calculated in the same way. The synthetic sample was placed in the given bin. This process is repeated until the number of training subjects reaches 10 times of the original training dataset.

Model Validation

To evaluate our proposed model, we utilized fivefold cross-validation for both classification and regression tasks. Specifically, we randomly divided the dataset into five portions. While one portion was used for testing, the remaining four portions were used as training data (70% for model training and 30% for model validation). This process was repeated five times until all portions of the dataset were treated as testing data. We evaluated the performance of risk prediction using accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC) across the five iterations. For cognitive score regression, we used Pearson's correlation coefficient, MAE, and standard deviation of absolute error (STD of AE). The fivefold cross-validation experiment was repeated 50 times to reduce the variability and the 95% confidence interval was reported. All the experiments are performed on a Windows 10 workstation with Intel Xeon Silver 4116 CPU @ 2.10 GHz, 128 GB RAM, and dual GTX 1080ti GPUs.

Most Discriminative Features Detection

In addition to the prediction of cognitive deficit, we seek to identify which brain regions contributed most to discriminate cognitive deficit. We used gradient-weighted class activation mapping (Grad-CAM) to highlight discriminative edges in the brain structural connectome map (Selvaraju et al., 2017). The Grad-CAM produces a coarse localization map highlighting predictive regions in the adjacency matrix by using gradient information of the last convolutional layer of the TL-CNN.

RESULTS

Subjects

After excluding the data with large motion artifacts, we had a total of 80 very preterm infants out of 110 subjects in the final analysis. The 80 subjects had a mean (SD) gestational age at birth of 28.0 (2.4) weeks and postmenstrual age at the scan of 40.4 (0.6) weeks. There are 41 (51.3%) male subjects. The mean (SD) birth weight of the cohort was 1091.5 (385.3) g. We considered the infants with Bayley III cognitive scores <90 as a high-risk group (31 subjects) and with Bayley III cognitive scores ≥ 90 as a low-risk group (49 subjects) to develop later moderate/severe cognitive deficits (Spencer-Smith et al., 2015).

Performance on Risk Stratification of Cognitive Deficits

We compared the proposed TL-CNN model with LR, linear SVM, and TL-DNN in the identification of very preterm infants at high-risk for moderate/severe cognitive deficits (**Table 1**). The receiver operating characteristic curves of various machine learning models are displayed in **Figure 2**. Our proposed TL-CNN model achieved the best prediction performance among the compared models, with 74.5% on the balanced accuracy, 78.7% on specificity, 70.2% on sensitivity, and 0.75 on AUC. The CNN model achieved the lowest balanced accuracy of 67.3%, while DNN had the lowest AUC of 0.59. We also noted that the linear SVM model had better AUC than both DNN and CNN.

Without the TL strategy, the CNN model achieved better accuracy and AUC than DNN. A similar trend was observed on CNN and DNN models with the TL strategy. The TL-DNN achieved 71.6% on the balanced accuracy, 76.8% on specificity, 66.4% on sensitivity, and 0.72 on AUC. The proposed TL-CNN model significantly improved the cognitive deficit prediction over the TL-DNN model by 2.9% in accuracy ($p = 0.005$) and 3.0% in AUC ($p = 0.008$). This demonstrated the advantage of treating brain structural connectome as images instead of vectorized weights.

Transfer learning-enhanced models (i.e., TL-DNN and TL-CNN) had significantly better prediction performance than the models without TL (i.e., DNN and CNN). TL strategy significantly improved prediction accuracy and AUC of CNN by 7.2% ($p < 0.001$) and 11.6% ($p < 0.001$). Similarly, TL-DNN increased prediction accuracy and AUC of DNN by 2.9% ($p = 0.002$) and 3.5% ($p < 0.001$). These results illustrated the effectiveness of the TL approach in deep learning models on the prediction of cognitive deficit.

Performance on the Prediction of Cognitive Scores

In the regression task, the proposed TL-CNN model had the highest Pearson's correlation coefficient ($r = 0.47$, $p < 0.001$) between the predicted and actual cognitive scores compared to linear regression ($r = 0.29$, $p < 0.001$), support vector regression (SVR) ($r = 0.32$, $p < 0.001$), and TL-DNN ($r = 0.37$, $p < 0.001$) models (**Table 2**). TL-CNN had the lowest mean STD of AE of 9.5.

Discriminative Brain Structural Connectome

To reveal which brain regions contributed to the prediction of cognitive deficits, we identified the predictive brain structural connections using the Grad-CAM method (Selvaraju et al., 2017). **Table 3** displays the top 15 predictive brain structural connections. We further demonstrated the identified brain connections in a circos plot (**Figure 3**). The top three discriminative structural connections are located within frontal lobes, limbic lobes, and the subcortical structure. We also plotted those discriminative connections on a brain atlas region using BrainNet Viewer (Xia et al., 2013; **Supplementary Figure 2**).

DISCUSSION

Early diagnosis and prediction of cognitive deficit for very preterm infants remain very challenging yet critical for early intervention. In this study, we proposed a TL-CNN model using brain structural connectome at term-equivalent age to predict future cognitive outcomes (i.e., standardized Bayley III cognitive scores). The TL-CNN model achieved promising performance in both risk classification and score regression tasks. For risk prediction of cognitive deficit, TL-CNN achieved a balanced

TABLE 1 | Performance of various machine learning models in utilizing the structural connectome at term-equivalent age to predict cognitive deficits at 2 years corrected age in very preterm infants.

Models	Balanced accuracy (%)	Specificity (%)	Sensitivity (%)	AUC
LR	68.3 (67.5, 72.0)	72.3 (71.2, 73.8)	64.4 (62.4, 66.5)	0.65 (0.63, 0.67)
SVM	70.5 (67.7, 71.7)	76.9 (74.8, 78.9)	64.0 (61.8, 66.1)	0.69 (0.67, 0.71)
DNN	68.7 (65.7, 69.5)	75.0 (72.9, 77.1)	62.5 (60.4, 64.5)	0.59 (0.57, 0.61)
CNN	67.3 (66.2, 70.2)	73.7 (71.7, 75.6)	61.0 (59.1, 62.9)	0.64 (0.62, 0.73)
TL-DNN	71.6 (70.7, 73.1)	76.8 (75.8, 77.9)	66.4 (65.0, 67.7)	0.72 (0.70, 0.74)
TL-CNN	**74.5 (73.4, 76.0)**	**78.7 (77.2, 79.8)**	**70.2 (68.5, 70.7)**	**0.75 (0.74, 0.76)**

Data in brackets are 95% confidence intervals. LR, logistic regression; SVM, support vector machine; TL-DNN, transfer learning enhanced deep neural network; TL-CNN, transfer learning enhanced convolutional neural network; AUC, area under the receiver operating characteristic curve.

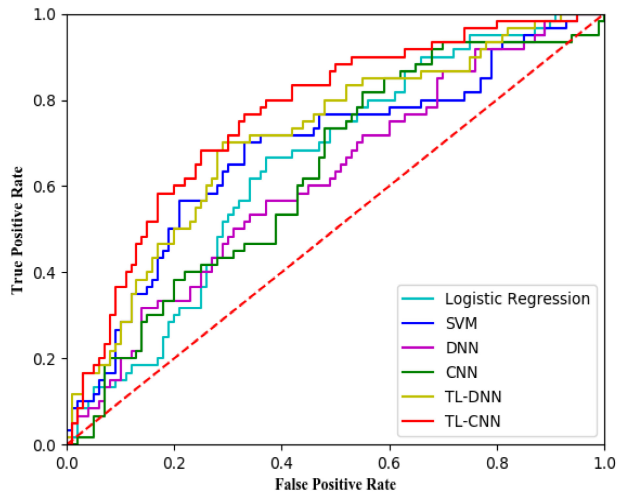


FIGURE 2 | Receiver operating characteristic (ROC) curves of different prediction models using structural brain connectome at term-equivalent age in predicting cognitive deficits at 2 years corrected age in very preterm infants. The proposed TL-CNN model achieved the best area under the ROC curve among compared machine learning models. SVM, support vector machine; DNN, deep neural network; CNN, convolutional neural network; TL-DNN, transfer learning enhanced deep neural network; TL-CNN, transfer learning-enhanced convolutional neural network.

accuracy of 74.5% and an AUC of 0.75. For regression of cognitive scores, the TL-CNN model had the best Pearson's correlation coefficient among multiple machine learning models. These results demonstrated the feasibility and advantages of a deep learning model that may facilitate the early diagnosis and classification of cognitive deficit for very preterm infants at term-equivalent age.

The proposed TL-CNN model outperformed several peer machine learning models by using both spatial and topological locality information embedded in the adjacency matrix of the brain structural connectome. For those traditional and fully connected neuron-based DNN, the brain connectome is flattened to a vector (He et al., 2018; Chen et al., 2019). This approach discards important spatial and topological locality information from the connectome. By treating the brain structural connectome as 2D images, convolutional filters in CNN can inherently learn the spatial information. In this study, we adopted a neonatal AAL brain atlas (Shi et al., 2011). The regions in this atlas are numbered through 1–90 and spatially nearby regions have adjacent numbers. In the adjacency matrix of structural connectome, the location of the brain regions follows the original ordering of 1–90; therefore, brain regions near each other in the structural connectome are typically near each other in Euclidean/brain space. In this way, convolutional filters of CNN are able to learn spatial connectivity information of those "clustered" nearby regions. Meanwhile, those 2D grid convolutional filters move in both row and column directions across the whole adjacency matrix in a single convolutional layer. After a series of consecutive layers, the deep CNN model can integrate the topological locality information gradually. Thus, we believe that applying the deep CNN model on the adjacency matrix provides unique insight to learn latent spatial and topological locality embedded in the brain structural connectome. The significantly improved prediction performance by the proposed TL-CNN supports the rationale of our study design.

We applied CNN to learn the spatial and topological information of the structural connectome. In this study, we constructed the structural connectome based on a neonatal AAL brain atlas (Shi et al., 2011). The regions in this atlas are numbered through 1–90 and spatially nearby regions have adjacent numbers. Specifically, the neonatal AAL atlas arranged 90 brain regions into the following sections: frontal lobe (region: 1–28, 69–70), occipital lobe (region: 43–54), parietal lobe (region: 61–68), central structures (region: 55–60), and temporal lobe (region: 37–42, 71–90). Therefore, though not strictly speaking, brain regions near each other in the structural connectome are typically near each other in Euclidean/brain space. As CNN's convolutional filters move across rows and columns of the structural connectome adjacency matrix in a moving-windows manner, the model was able to learn topological connectivity information. We tested the prediction performance with five different permuted connectome matrices. The TL-CNN achieved an accuracy of 68.8% (95% CI, 66.9, 70.7), and an AUC of 0.65 (95% CI, 0.63, 0.67), which was slightly lower than the performance of using original structure connectome matrix. This indicates that the order of the ROIs in the structural connectome matrix has an impact on the outcome prediction performance.

Transfer learning technique is essential for studies of very preterm infants using deep learning models. The big data revolution has boosted recent advances in deep learning techniques. Without large training samples, it is very difficult to train a complex deep learning model from scratch. Indeed, the linear SVM demonstrated better performance than deep learning models without the TL strategy in our study. Deep learning models trained with a small number of samples tend to be overfitted. Those relatively simple machine learning models (e.g., SVM) may achieve better performance. Unfortunately, the availability of annotated large brain imaging datasets with clinical and outcome information from very preterm infants is usually very limited, preventing the application of deep learning models in this research domain. The CNN model is a complex network consisting of millions of trainable weights that requires a large amount of data to update the weights when training the model. The TL technique addressed this issue by applying knowledge learned from a large dataset in the source domain to a new target task with limited data to improve performance and model robustness. In the present study, we transferred the knowledge (i.e., optimized weights) from a pre-trained model to the prediction/regression tasks in the target domain and then fine-tuned the model using brain structural connectome to optimize the performance of risk prediction/score regression. The increased performance supports our hypothesis regarding the effectiveness of the TL strategy.

The data balance and augmentation technique also improved the model training. Our dataset was imbalanced with a small number of subjects having low Bayley III cognitive scores. The

TABLE 2 | Performance of various machine learning models in utilizing the structural connectome at term-equivalent age to predict Bayley-III cognitive scores at 2 years corrected age in very preterm infants.

Models	*r*	*p*	MAE	STD of AE
Linear regression	0.29 (0.27, 0.31)	<0.0001	20.1 (17.6, 22.6)	12.0 (10.7, 13.3)
SVR	0.32 (0.31, 0.34)	<0.0001	18.2 (15.1, 20.9)	11.4 (9.4, 13.4)
TL-DNN	0.37 (0.35, 0.39)	<0.0001	22.5 (20.0, 24.9)	11.2 (9.5, 13.0)
TL-CNN	**0.47 (0.45, 0.49)**	**<0.0001**	**16.2 (13.8, 18.5)**	**9.5 (7.8, 11.2)**

Data in brackets are 95% confidence intervals. r, correlation between true and predicted Bayley-III cognitive scores; p, p-value (false discovery rate corrected) of one-sample t-test of r; MAE, mean absolute error; STD of AE, standard deviation of absolute error; SVR, support vector regression; TL-DNN, transfer learning enhanced deep neural network; TL-CNN, transfer learning enhanced convolutional neural network.

TABLE 3 | Top 15 discriminative brain structural connections for prediction of cognitive deficits.

Brain region A	Abbreviation	Brain region B	Abbreviation	*r*
Top discriminative features				
Precentral gyrus left	PreCG-L	Putamen left	PUT-L	0.39
Superior occipital gyrus left	SOG-L	Superior occipital gyrus right	SOG-R	0.37
Hippocampus left	HIP-L	Middle occipital gyrus left	MOG-L	0.34
Postcentral gyrus right	PoCG-R	Putamen right	PUT-R	0.33
Hippocampus right	HIP-R	Postcentral gyrus right	PoCG-R	0.33
Hippocampus left	HIP-L	Superior parietal gyrus left	SPG-L	0.33
Orbitofrontal cortex (superior) left	ORBsup-L	Orbitofrontal cortex (medial) right	ORBmed-R	0.29
Putamen left	PUT-L	Hippocampus left	HIP-L	0.28
Postcentral gyrus left	PoCG-L	Putamen left	PUT-L	0.27
Putamen right	PUT-R	Hippocampus right	HIP-R	−0.25
Postcentral gyrus left	PoCG-L	Hippocampus left	HIP-L	0.25
Hippocampus right	HIP-R	Thalamus right	THA-R	0.21
Cuneus left	CUN-L	Precuneus right	PCUN-R	−0.21
Cuneus left	CUN-L	Superior occipital gyrus right	SOG-R	0.20
Superior frontal gyrus (dorsal) right	SFGdor-R	Hippocampus right	HIP-R	0.19

r, correlation between brain connectome weights and true Bayley-III cognitive scores.

imbalanced dataset may result in a model that is more likely to predict a high-risk subject into the majority low-risk group. Thus, we applied the data balance and augmentation technique before training any model in this work.

Identification of discriminative brain regions not only improves our understanding of the neurodevelopment of very preterm infants but also enhances the integrity of trained deep learning models. We applied the Grad-CAM method to rank the importance of individual links. Multiple brain regions such as postcentral gyrus, thalamus, and superior occipital gyrus were identified by our TL-CNN model to be predictive to cognitive deficits. These regions were also found to be predictive in our previous study using functional connectome on an independent cohort (He et al., 2018). In addition, postcentral gyrus, thalamus, and superior occipital gyrus were also reported in prior independent studies (Corbetta, 1998; Ouhaz et al., 2018), indicating their association with brain cognitive function. These somatosensory regions are thought to be part of the mirror system, which plays an important role in imitating, understanding, and learning for brain cognitive development (Acharya and Shukla, 2012). Furthermore, the identified most predictive regions have been associated with emotional regulation and memory (limbic lobe) (Catani et al., 2013), visual processing (occipital lobe) (Pöppel et al., 1978), and sensory, visual, and language information processing (parietal lobe) (Wolpert et al., 1998). Additionally, subcortical gray matter regions that play an important role in motion preparation and execution were also ranked highly by the proposed TL-CNN model (Chang et al., 2018).

We further performed a correlation analysis between the top 15 discriminative structural connectome connections and the cognitive outcomes at 2 years corrected age (**Table 3**). Briefly, the majority of brain connectome connections have a positive correlation with the cognitive scores. The increased connectivity strength of these connections would indicate a lower risk of cognitive deficits in very preterm infants at 2 years corrected age. This trend is consistent with our previous study (He and Parikh, 2016). In contrast, two brain connectome connections (Putamen right–Hippocampus right and Cuneus left–Precuneus right) are negatively correlated with cognitive scores, indicating that the increased connectivity strength of these two connections suggests a higher risk of cognitive deficits in very preterm infants at 2 years corrected age. Further investigation is required to unveil the underlying pathological mechanism of these brain connectome connections on brain cognitive functions.

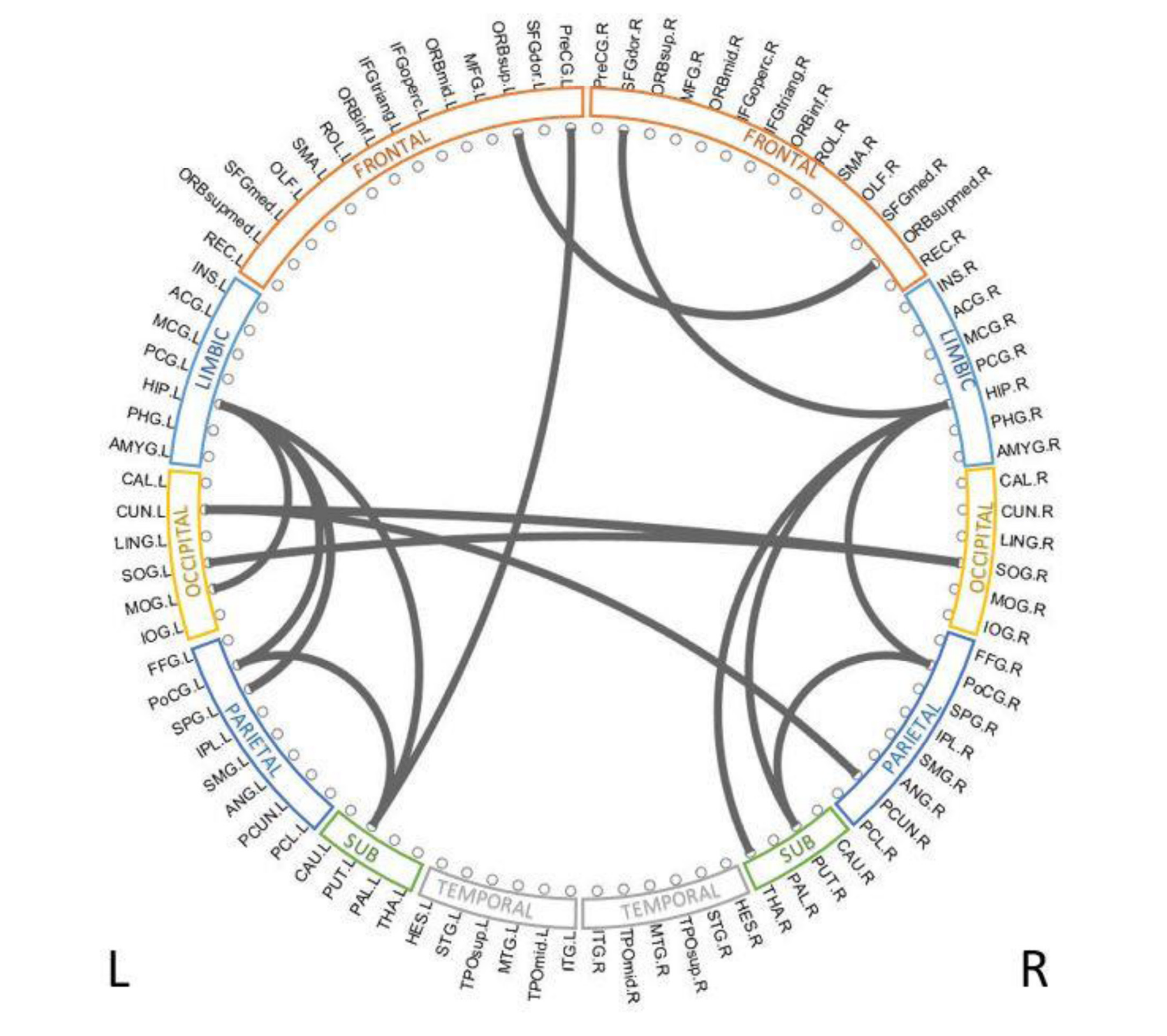


FIGURE 3 | Top 15 discriminative brain structural connections identified by TL-CNN, a circos plot. The top three discriminative structural connections are located within frontal lobes, limbic lobes, and the subcortical structure.

There are several limitations to this study. First, we only internally validated our data in our cohort of very preterm infants. External datasets from independent studies or other research groups are necessary to externally validate the proposed TL-CNN models. Second, we only used brain structural connectome data for the outcome prediction. The integration of brain functional connectome and/or clinical data in our model is likely to improve prediction performance. Third, we constructed brain structural connectome based on an AAL brain atlas without cerebellum regions (Shi et al., 2011). However, the cerebellum regions have been conventionally recognized to have an impact on motor function and recently have been proven to associate with cognitive function (Schmahmann, 2019). The inclusion of the cerebellum regions when we construct the structural connectome may further enhance the prediction performance.

CONCLUSION

In summary, this study proposed a deep learning model TL-CNN for early prediction of cognitive deficit in very preterm infants at 2 years corrected age using brain structural connectome derived from DTI obtained at term-equivalent age. The proposed model achieved improved performance by integrating multiple technique advances, including the convolution of CNN on adjacency matrix, TL strategy, and data balance and augmentation approach. The results suggest that deep learning models may facilitate early prediction of later neurodevelopmental outcomes in very preterm infants at term-equivalent age.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board Nationwide Children's Hospital. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MC, HL, NP, and LH designed the study. MC, HL, JW, and LH conducted the experiments and data analysis. All authors wrote the manuscript.

FUNDING

This study was supported by the National Institutes of Health grants R21-HD094085, R01-NS094200, and R01-NS096037 and a Trustee grant from Cincinnati Children's Hospital Medical Center.

ACKNOWLEDGMENTS

We sincerely thank Jennifer Notestine, RN, Valerie Marburger, NNP, Julie Gutentag, RN, and Patty Luzader, RN, for serving as the study coordinators and Mark Smith, MS, for serving as the study MR tech. We are grateful to the families and neonatal intensive care unit and clinic staff that made this study possible.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00858/full#supplementary-material

REFERENCES

Acharya, S., and Shukla, S. (2012). Mirror neurons: Enigma of the metaphysical modular brain. *J. Nat. Sci. Biol. Med.* 3, 118–124. doi: 10.4103/0976-9668.101878

Azizi, S., Mousavi, P., Yan, P., Tahmasebi, A., Kwak, J. T., Xu, S., et al. (2017). Transfer learning from RF to B-mode temporal enhanced ultrasound features for prostate cancer detection. *Int. J. Comput. Assist. Radiol. Surg.* 12, 1111–1121. doi: 10.1007/s11548-017-1573-x

Barkhof, F., Haller, S., and Rombouts, S. A. (2014). Resting-state functional MR imaging: a new window to the brain. *Radiology* 272, 29–49. doi: 10.1148/radiol.14132388

Bassett, D. S., Brown, J. A., Deshpande, V., Carlson, J. M., and Grafton, S. T. (2011). Conserved and variable architecture of human white matter connectivity. *Neuroimage* 54, 1262–1279. doi: 10.1016/j.neuroimage.2010.09.006

Bengio, Y. (2012). "Deep Learning of Representations for Unsupervised and Transfer Learning," in *Paper presented at the Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, (Berlin: Springer).

Bizzego, A., Bussola, N., Chierici, M., Maggio, V., Francescatto, M., Cima, L., et al. (2019). Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. *PLoS Comput. Biol.* 15:e1006269. doi: 10.1371/journal.pcbi.1006269

Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A. B., Narwal, R., et al. (2012). National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *Lancet* 379, 2162–2172. doi: 10.1016/S0140-6736(12)60820-4

Bullmore, E., and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.* 10, 186–198. doi: 10.1038/nrn2575

Catani, M., Dell'acqua, F., and Thiebaut de Schotten, M. (2013). A revised limbic system model for memory, emotion and behaviour. *Neurosci. Biobehav. Rev.* 37, 1724–1737. doi: 10.1016/j.neubiorev.2013.07.001

Chang, D. H. F., Ban, H., Ikegaya, Y., Fujita, I., and Troje, N. F. (2018). Cortical and subcortical responses to biological motion. *Neuroimage* 174, 87–96. doi: 10.1016/j.neuroimage.2018.03.013

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *J. Ar. Intell. Res.* 16, 321–357. doi: 10.1613/jair.953

Chen, M., Li, H., Wang, J., Dillman, J. R., Parikh, N. A., and He, L. (2019). A Multichannel Deep Neural Network Model Analyzing Multiscale Functional Brain Connectome Data for Attention Deficit Hyperactivity Disorder Detection. *Radiol. Ar. Intell.* 2:e190012. doi: 10.1148/ryai.2019190012

Cheng, B., Liu, M. X., Zhang, D. Q., Shen, D. G., and Alzheimer's Disease Neuroimaging Initiative (2019). Robust multi-label transfer feature learning for early diagnosis of Alzheimer's disease. *Brain Imaging Behav.* 13, 138–153. doi: 10.1007/s11682-018-9846-8

Choi, J. Y., Yoo, T. K., Seo, J. G., Kwak, J., Um, T. T., and Rim, T. H. (2017). Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database. *PLoS One* 12:e0187336. doi: 10.1371/journal.pone.0187336

Corbetta, M. (1998). Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proc. Nat. Acad. Sci. Unit. Stat. Am.* 95, 831–838. doi: 10.1073/pnas.95.3.831

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., and Li, F. F. (2009). "ImageNet: A Large-Scale Hierarchical Image Database," in *Cvpr: 2009 Ieee Conference on Computer Vision and Pattern Recognition*, (New Jersey: IEEE).

Fortin, J. P., Parker, D., Tunc, B., Watanabe, T., Elliott, M. A., Ruparel, K., et al. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* 161, 149–170. doi: 10.1016/j.neuroimage.2017.08.047

Gao, Y. F., Zhang, Y. M., Wang, H. L., Guo, X. J., and Zhang, J. C. (2019). Decoding Behavior Tasks From Brain Activity Using Deep Transfer Learning. *IEEE Access* 7, 43222–43232. doi: 10.1109/Access.2019.2907040

Girault, J. B., Munsell, B. C., Puechmaille, D., Goldman, B. D., Prieto, J. C., Styner, M., et al. (2019). White matter connectomes at birth accurately predict cognitive abilities at age 2. *NeuroImage* 192, 145–155. doi: 10.1016/j.neuroimage.2019.02.060

Gupta, A., Ayhan, M. S., and Maida, A. S. (2013). *Natural image bases to represent neuroimaging data*. Atlanta, GA: JMLR.org.

Hamilton, B. E., Martin, J. A., and Osterman, M. J. (2016). Births: Preliminary Data for 2015. *Natl. Vital. Stat. Rep.* 65, 1–15.

He, L., Li, H., Holland, S. K., Yuan, W., Altaye, M., and Parikh, N. A. (2018). Early prediction of cognitive deficits in very preterm infants using functional connectome data in an artificial neural network framework. *Neuroim. Clin.* 18, 290–297. doi: 10.1016/j.nicl.2018.01.032

He, L., and Parikh, N. A. (2016). Brain functional network connectivity development in very preterm infants: The first six months. *Early Hum. Dev.* 98, 29–35. doi: 10.1016/j.earlhumdev.2016.06.002

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImag. Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Hess, C. P., Mukherjee, P., Han, E. T., Xu, D., and Vigneron, D. B. (2006). Q-ball reconstruction of multimodal fiber orientations using the spherical harmonic basis. *Magnet. Reson. Med.* 56, 104–117. doi: 10.1002/mrm.20931

Kawahara, J., Brown, C. J., Miller, S. P., Booth, B. G., Chau, V., Grunau, R. E., et al. (2017). BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *Neuroimage* 146, 1038–1049. doi: 10.1016/j.neuroimage.2016.09.046

Kingma, D. P., and Ba, J. J. (2014). Adam: A method for stochastic optimization. *arXiv*

Kooi, T., van Ginneken, B., Karssemeijer, N., and den Heeten, A. (2017). Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Med. Phys.* 44, 1017–1027. doi: 10.1002/mp.12110

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

Li, H. L., Parikh, N. A., and He, L. L. (2018). A Novel Transfer Learning Approach to Enhance Deep Neural Network Classification of Brain Functional Connectomes. *Front. Neurosci.* 12:491. doi: 10.3389/fnins.2018.00491

Munsell, B. C., Wee, C. Y., Keller, S. S., Weber, B., Elger, C., da Silva, L. A. T., et al. (2015). Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *Neuroimage* 118, 219–230. doi: 10.1016/j.neuroimage.2015.06.008

Nagy, Z., Westerberg, H., and Klingberg, T. (2004). Maturation of white matter is associated with the development of cognitive functions during childhood. *J. Cogn. Neurosci.* 16, 1227–1233. doi: 10.1162/0898929041920441

Ouhaz, Z., Fleming, H., and Mitchell, A. S. (2018). Cognitive Functions and Neurodevelopmental Disorders Involving the Prefrontal Cortex and Mediodorsal Thalamus. *Front. Neurosci.* 12:3310. doi: 10.3389/fnins.2018.00033

Park, H.-J., and Friston, K. J. S. (2013). Structural and functional brain networks: from connections to cognition. *Science* 342:1238411. doi: 10.1126/science.1238411

Petersen, S. E., and Sporns, O. J. N. (2015). Brain networks and cognitive architectures. *Neuron* 88, 207–219. doi: 10.1016/j.neuron.2015.09.027

Pöppel, E., Brinkmann, R., von Cramon, D., and Singer, W. (1978). Association and dissociation of visual functions in a case of bilateral occipital lobe infarction. *Arch. Psyc. Nervenkrankheiten* 225, 1–21. doi: 10.1007/bf00367348

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). *Self-taught learning: transfer learning from unlabeled data*. Oregon, USA: Association for Computing Machinery.

Samala, R. K., Chan, H. P., Hadjiiski, L., Helvie, M. A., Wei, J., and Cha, K. (2016). Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Med. Phys.* 43:6654. doi: 10.1118/1.4967345

Samala, R. K., Chan, H. P., Hadjiiski, L. M., Helvie, M. A., Richter, C., and Cha, K. (2018). Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Phys. Med. Biol.* 63:095005. doi: 10.1088/1361-6560/aabb5b

Schmahmann, J. D. (2019). The cerebellum and cognition. *Neurosci. Lett.* 688, 62–75. doi: 10.1016/j.neulet.2018.07.005

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Paper presented at the Proceedings of the IEEE international conference on computer vision*, (New Jersey: IEEE).

Sen, B., Borle, N. C., Greiner, R., and Brown, M. R. G. (2018). A general prediction model for the detection of ADHD and Autism using structural and functional MRI. *PLoS One* 13:e0194856. doi: 10.1371/journal.pone.0194856

Shi, F., Yap, P. T., Wu, G. R., Jia, H. J., Gilmore, J. H., Lin, W. L., et al. (2011). Infant Brain Atlases from Neonates to 1-and 2-Year-Olds. *PLoS One* 6:e1874610. doi: 10.1371/journal.pone.0018746

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Trans. Med, Imag.* 35, 1285–1298. doi: 10.1109/TMI.2016.2528162

Simonyan, K., and Zisserman, A. J. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*

Spencer-Smith, M. M., Spittle, A. J., Lee, K. J., Doyle, L. W., and Anderson, P. J. (2015). Bayley-III Cognitive and Language Scales in Preterm Children. *Pediatrics* 135, e1258–e1265. doi: 10.1542/peds.2014-3039

Sporns, O., Tononi, G., and Kötter, R. (2005). The human connectome: a structural description of the human brain. *PLoS comput. Biol.* 1:e42. doi: 10.1371/journal.pcbi.0010042

Wang, R., Benner, T., Sorensen, A. G., and Wedeen, V. J. (2007). Diffusion toolkit: a software package for diffusion imaging data processing and tractography. *Proc. Intl. Soc. Mag. Reson Med.* 15:3720.

Wang, Y., Wang, C., and Zhang, H. (2018). Ship Classification in High-Resolution SAR Images Using Deep Learning of Small Datasets. *Sensors* 18:2929. doi: 10.3390/s18092929

Wee, C. Y., Yap, P. T., Zhang, D., Denny, K., Browndyke, J. N., Potter, G. G., et al. (2012). Identification of MCI individuals using structural and functional connectivity networks. *Neuroimage* 59, 2045–2056. doi: 10.1016/j.neuroimage.2011.10.015

Wolpert, D. M., Goodbody, S. J., and Husain, M. (1998). Maintaining internal representations: the role of the human superior parietal lobe. *Nat. Neurosci.* 1, 529–533. doi: 10.1038/2245

Xia, M. R., Wang, J. H., and He, Y. (2013). BrainNet Viewer: A Network Visualization Tool for Human Brain Connectomics. *PLoS One* 8:e6891010. doi: 10.1371/journal.pone.0068910

Yuan, W. H., Wade, S. L., and Babcock, L. (2015). Structural Connectivity Abnormality in Children with Acute Mild Traumatic Brain Injury using Graph Theoretical Analysis. *Hum. Brain Map.* 36, 779–792. doi: 10.1002/hbm.22664

Zhen, X., Chen, J., Zhong, Z., Hrycushko, B., Zhou, L., Jiang, S., et al. (2017). Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. *Phys. Med. Biol.* 62, 8246–8263. doi: 10.1088/1361-6560/aa8d09

Zheng, J., Miao, S., Jane Wang, Z., and Liao, R. (2018). Pairwise domain adaptation module for CNN-based 2-D/3-D registration. *J. Med. Imaging* 5:021204. doi: 10.1117/1.JMI.5.2.021204

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 Neuroscience

REVIEW
published: 08 October 2020
doi: 10.3389/fnins.2020.00779

A Survey on Deep Learning for Neuroimaging-Based Brain Disorder Analysis

Li Zhang[1,2]*, Mingliang Wang*[2]*, Mingxia Liu*[3]** and Daoqiang Zhang*[2]*

[1] College of Computer Science and Technology, Nanjing Forestry University, Nanjing, China, [2] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China, [3] Department of Radiology and Biomedical Research Imaging Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States

Deep learning has recently been used for the analysis of neuroimages, such as structural magnetic resonance imaging (MRI), functional MRI, and positron emission tomography (PET), and it has achieved significant performance improvements over traditional machine learning in computer-aided diagnosis of brain disorders. This paper reviews the applications of deep learning methods for neuroimaging-based brain disorder analysis. We first provide a comprehensive overview of deep learning techniques and popular network architectures by introducing various types of deep neural networks and recent developments. We then review deep learning methods for computer-aided analysis of four typical brain disorders, including Alzheimer's disease, Parkinson's disease, Autism spectrum disorder, and Schizophrenia, where the first two diseases are neurodegenerative disorders and the last two are neurodevelopmental and psychiatric disorders, respectively. More importantly, we discuss the limitations of existing studies and present possible future directions.

Keywords: deep learning, neuroimage, Alzheimer's disease, Parkinson's disease, autism spectrum disorder, schizophrenia

Edited by:
Jun Shi,
Shanghai University, China

Reviewed by:
Xia Wu,
Beijing Normal University, China
Jinglei Lv,
The University of Sydney, Australia

Correspondence:
Mingxia Liu
mxliu@med.unc.edu
Daoqiang Zhang
dqzhang@nuaa.edu.cn

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *10 May 2020*
Accepted: *02 June 2020*
Published: *08 October 2020*

Citation:
Zhang L, Wang M, Liu M and Zhang D (2020) A Survey on Deep Learning for Neuroimaging-Based Brain Disorder Analysis. Front. Neurosci. 14:779. doi: 10.3389/fnins.2020.00779

1. INTRODUCTION

Medical imaging refers to several different technologies that are used to provide visual representations of the interior of the human body in order to aid the radiologists and clinicians to detect, diagnose, or treat diseases early and more efficiently (Brody, 2013). Over the past few decades, medical imaging has quickly become a dominant and effective tool and represents various imaging modalities, including X-ray, mammography, ultrasound, computed tomography, magnetic resonance imaging (MRI), and positron emission tomography(PET) (Heidenreich et al., 2002). Each type of these technologies gives various pieces of anatomical and functional information about the different body organs for diagnosis as well as for research. In clinical practice, the detail interpretation of medical images needs to be performed by human experts, such as the radiologists and clinicians. However, for the enormous number of medical images, the interpretations are time-consuming and easily influenced by the biases and potential fatigue of human experts. Therefore, from the early 1980s, doctors and researchers have begun to use computer-assisted diagnosis (CAD) systems to interpret the medical images and to improve their efficiency.

In the CAD systems, machine learning is able to extract informative features that describe the inherent patterns from data and play a vital role in medical image analysis (Wernick et al., 2010; Wu et al., 2016; Erickson et al., 2017; Li et al., 2019). However, the structures of the medical images are very complex, and the feature selection step is still carried out by the human experts on the basis of their domain-specific knowledge. This results in a challenge for non-experts to utilize machine learning techniques in medical image analysis. Therefore, the handcrafted feature selection is not suitable for medical images. Though the sparse learning and dictionary learning have demonstrated the validity of these techniques for automatically discovering discriminative features from training samples, the shallow architectures of these algorithms limit their representational power (Pandya et al., 2019).

Compared to the traditional machine learning algorithms, deep learning automatically discovers the informative representations without the professional knowledge of domain experts and allows the non-experts to effectively use deep learning techniques. Therefore, deep learning has rapidly becomes a methodology of choice for medical image analysis in recent years (LeCun et al., 2015; Schmidhuber, 2015; Goodfellow et al., 2016; Lian et al., 2018). Due to enhanced computer power with the high-tech central processing units (CPU) and graphical processing units (GPU), the availability of big data, and the creation of novel algorithms to train deep neural networks, deep learning has seen unprecedented success in the most artificial intelligence applications, such as computer vision (Voulodimos et al., 2018), natural language processing (Sarikaya et al., 2014), and speech recognition (Bahdanau et al., 2016). Especially, the improvement and successes of computer vision simultaneously prompted the use of deep learning in the medical image analysis (Lee et al., 2017; Shen et al., 2017).

Currently, deep learning has fueled great strides in medical image analysis. We can divide the medical image analysis tasks into several major categories: classification, detection/localization, registration, and segmentation (Litjens et al., 2017). The classification is one of the first tasks in which deep learning giving a major contribution to medical image analysis. This task aims to classify medical images into two or more classes. The stacked auto-encoder model was used to identify Alzheimer's disease or mild cognitive impairment by combining medical images and biological features (Suk et al., 2015). The detection/localization task consists of the localization and identification of the landmarks or lesion in the full medical image. For example, deep convolutional neural networks were used for the detection of lymph nodes in CT images (Roth et al., 2014). The segmentation task is to partition a medical image into different meaningful segments, such as different tissue classes, organs, pathologies, or other biologically relevant structures (Sun et al., 2019a). The U-net was the most well-known deep learning architecture, which used convolutional networks for biomedical image segmentation (Ronneberger et al., 2015). Registration of medical images is a process that searches for the correct alignment of images. Wu et al. (2013) utilized convolutional layers to extract features from input patches in an unsupervised manner. Then the obtained feature vectors were used to replace the handcrafted features in the HAMMER registration algorithm. In addition, the medical image analysis contains other meaningful tasks, such as content-based image retrieval (Li et al., 2018c) and image generation and enhancement (Oktay et al., 2016) in combination with image data and reports (Schlegl et al., 2015).

There are many papers have comprehensively surveyed the medical image analysis using deep learning techniques (Lee et al., 2017; Litjens et al., 2017; Shen et al., 2017). However, these papers usually reviewed all human tissues, including the brain, chest, eye, breast, cardiac, abdomen, musculoskeletal, and others. Almost no papers focus on one specific tissue or disease (Hu et al., 2018). Brain disorders are among the most severe health problems facing our society, causing untold human suffering and enormous economic costs. Many studies successfully used medical imaging techniques for the early detection, diagnosis, and treatment of the human brain disorders, such as neurodegenerative disorders, neurodevelopmental disorders and psychiatric disorders (Vieira et al., 2017; Durstewitz et al., 2019). We therefore pay more close attention to human brain disorders in this survey. About 100 papers are reviewed, most of them published from 2016 to 2019, on deep learning for brain disorder analysis.

The structure of this review can roughly be divided into two parts, the deep learning architectures and the usage of deep learning in brain disorder analysis and is organized as follows. In section 2, we briefly introduce some popular deep learning models. In section 3, we provide a detailed overview of recent studies using deep learning techniques for four brain disorders, including Alzheimer's disease, Parkinson's disease, Autism spectrum disorder, and Schizophrenia. Finally, we analyze the limitations of the deep learning techniques in medical image analysis and provide some research directions for further study. For the convenience of readers, the abbreviations of terminologies used in the following context are listed in the **Supplementary Table 1**.

2. DEEP LEARNING

In this section, we introduce the fundamental concept of basic deep learning models in the literature, which have been wildly applied to medical image analysis, especially human brain disorder diagnosis. These models include feed-forward neural networks, deep generative models (e.g., stacked auto-encoders, deep belief networks, deep Boltzmann machine, and generative adversarial networks), convolutional neural networks, graph convolutional networks, and recurrent neural networks.

2.1. Feed-Forward Neural Networks

In machine learning, artificial neural networks (ANN) aim to simulate intelligent behavior by mimicking the way that biological neural networks function. The simplest artificial neural networks is a single-layer architecture, which is composed of an input layer and an output layer (**Figure 1A**). However, despite the use of non-linear activation functions in output layers, the single-layer neural network usually obtains poor performance for complicated data patterns. In order to circumvent the limitation, the multi-layer perceptron (MLP), also referred to as a feed-forward neural network (FFNN) (**Figure 1B**), which includes a so-call hidden layer between the input layer and the output layer. Each layer contains multiple units which are fully connected to

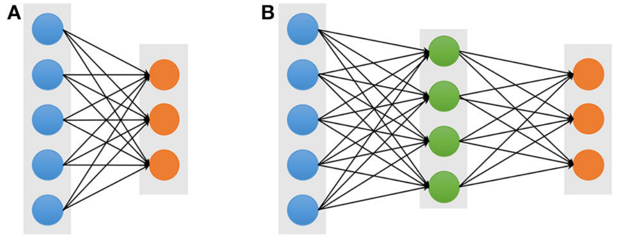


FIGURE 1 | Architectures of the single-layer **(A)** and multi-layer **(B)** neural networks. The blue, green, and orange solid circles represent the input visible, hidden, and output units, respectively.

units of neighboring layers, but there are no connections between units in the same layer. Given an input visible vector $\boldsymbol{x}$, the composition function of output unit $\boldsymbol{y_k}$ can be written as follows:

$$y_k(\boldsymbol{x};\boldsymbol{\theta}) = f^{(2)}\left(\sum_{j=1}^{M} w_{k,j}^{(2)} f^{(1)}\left(\sum_{i=1}^{N} w_{j,i}^{(1)} x_i + b_j^{(1)}\right) + b_k^{(2)}\right) \quad (1)$$

where the superscript represents a layer index, M is the number of hidden units, and b_j and b_k represent the bias of input and hidden layer, respectively. $f^{(1)}(\cdot)$ and $f^{(2)}(\cdot)$ denote the non-linear activation function, and the parameter set is $\boldsymbol{\theta} = \{\boldsymbol{w}_j^{(1)}, \boldsymbol{w}_k^{(2)}, b_j^{(1)}, b_k^{(2)}\}$. The back-propagation(BP) is an efficient algorithm to evaluate a gradient in the FFNN (Rumelhart et al., 1986). The BP algorithm is to propagate the error values from the output layer back to the input layer through the network. Once the gradient vector of all the layers is obtained, the parameters $\boldsymbol{\theta}$ can be updated. Until the loss function is converged or the predefined number of iterations is reached, the update process stops and the network gets the model parameters $\boldsymbol{\theta}$.

2.2. Stacked Auto-Encoders

An auto-encoder (AE), also known as an auto-associator, learns the latent representations of input data (called encode) in an unsupervised manner and then uses these representations to reconstruct output data (called decode). Due to the simple and shallow structure, the power representation of a typical AE is relatively limited. However, when multiple AEs are stacked to form a deep network, called stacked auto-encoders (SAE) (**Figure 2**), the representation power of an SAE can be obviously improved (Bengio et al., 2007). Because of the deep structural characteristic, the SAE is able to learn and discover more complicated patterns inherent in the input data. The lower layers can only learn simpler data patterns, while the higher layers are able to extract more complicated data patterns. In a word, the different layers of an SAE represent different levels of data information (Shen et al., 2017). In addition, various AE variations, denoising auto-encoders (DAE) (Vincent et al., 2008), sparse auto-encoders (sparse AE) (Poultney et al., 2007), and variational auto-encoders (VAE) (Kingma and Welling, 2013), have been proposed and also can be stacked as SAE, such as the stacked sparse AE (SSAE) (Shin et al., 2013). These extensions of auto-encoders not only can learn more useful latent representations but also improve the robustness.

To avoid the drawback of the BP algorithm, which can cause the gradient falling into a poor local optimum (Larochelle et al., 2009), the greedy layer-wise approach is considered to training parameters of an SAE (Hinton and Salakhutdinov, 2006). The important character of the greedy layer-wise is to pre-train each layer in turn. In other words, the output of the l-th hidden layers is used as input data for the $(l+1)$-th hidden layer. The process performs as pre-training, which is conducted in an unsupervised manner with a standard BP algorithm. The important advantage of the pre-training is able to increase the size of the training dataset using unlabeled samples.

2.3. Deep Belief Networks

A Deep Belief Network (DBN) stacks multiple restricted Bolztman machines (RBMs) for deep architecture construction (Hinton et al., 2006). A DBN has one visible layer and multiple hidden layers as shown in **Figure 3A**. The lower layers form directed generative models. However, the top two layers form the distribution of RBM, which is an undirected generative model. Therefore, given the visible units $\boldsymbol{v}$ and L hidden layers $\boldsymbol{h}^{(1)}, \boldsymbol{h}^{(2)}, \ldots, \boldsymbol{h}^{(L)}$, the joint distribution of DBN is defined:

$$P(\boldsymbol{v}, \boldsymbol{h}^{(1)}, \ldots, \boldsymbol{h}^{(L)}) = P(\boldsymbol{v}|\boldsymbol{h}^{(1)})\left(\prod_{l=1}^{L-2} P(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l+1)})\right)P(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)}) \quad (2)$$

where $P(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l+1)})$ represents the conditional distribution for the units of the hidden layer l given the units of the hidden layer $l+1$, and $P(\boldsymbol{h}^{(L-1)}, \boldsymbol{h}^{(L)})$ corresponds the joint distribution of the top hidden layers $L-1$ and L.

As for training a DBN, there are two steps, including pre-training and fine-tuning. In the pre-training step, the sDBN is trained by stacking RBMs layer by layer to find the parameter space. Each layer is trained as an RBM. Specifically, the l-th hidden layer is trained as an RBM using the observation data from output representation of the $(l-1)$-th hidden layer, and this repeats, training each layer until the we reach the top layer. After the pre-training is completed, the fine-tuning is performed to further optimize the network to search the optimum parameters. The wake-sleep algorithm and the standard BP algorithm are good at fine-tuning for generative and discriminative models, respectively (Hinton et al., 1995). For a practical application problem, the obtained parameters from the pre-training step are used to initiate a DNN, and then the deep model can be fine-tuned by a supervised learning algorithm like BP.

2.4. Deep Boltzmann Machine

A Deep Boltzmann Machine (DBM) is also constructed by stacking multiple RBMs as shown in **Figure 3B** (Salakhutdinov and Larochelle, 2010; Salakhutdinov, 2015). However, unlike the DBN, all the layers of the DBM form an entirely undirected model, and each variable within the hidden layers are mutually independent. Thus, the hidden layer l is conditioned on its two neighboring layer $l-1$ and $l+1$, and its probability distribution is $P(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l-1)}, \boldsymbol{h}^{(l+1)})$. Given the values of the neighboring layers,

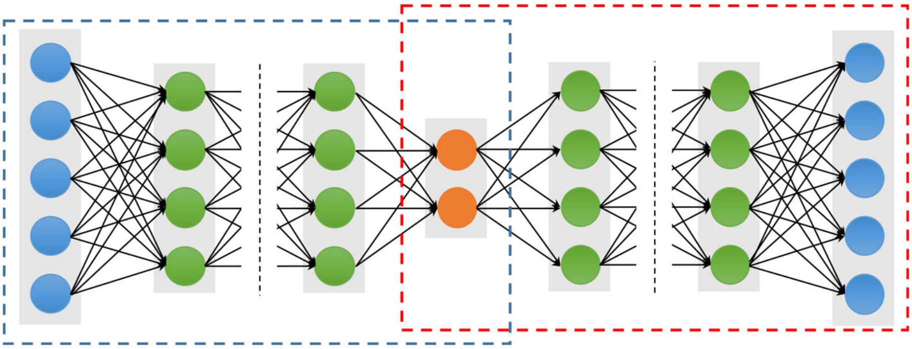


FIGURE 2 | Architectures of a stacked auto-encoder. The blue and red dotted boxes represent the encoding and decoding stage, respectively. The blue solid circles are the input and output units, which have the same number nodes. The orange solid circles represent the latent representation, and the green solid circles represent any hidden layers.

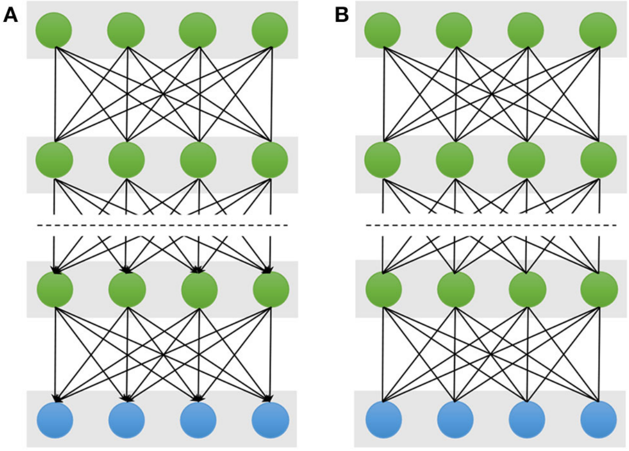


FIGURE 3 | Schematic illustration of Deep Belief Networks **(A)** and Deep Boltzmann Machine **(B)**. The double-headed arrow represents the undirected connection between the two neighboring layers, and the single-headed arrow is the directed connection. The top two layers of the DBN form an undirected generative model and the remaining layers form directed generative model. But all layers of the DBM are undirected generative model.

the conditional probabilities over the visible and the L set of hidden units are given by logistic sigmoid functions:

$$P(v_i|\boldsymbol{h}^1) = \sigma\big(\sum_j W_{ij}^{(1)} h_j^{(1)}\big) \tag{3}$$

$$P(h_k^{(l)}|\boldsymbol{h}^{(l-1)}, \boldsymbol{h}^{(l+1)}) = \sigma\big(\sum_m W_{mk}^{(l)} h_m^{(l-1)} + \sum_n W_{kn}^{(l+1)} h_n^{(l+1)}\big) \tag{4}$$

$$P(h_t^{(L)}|\boldsymbol{h}^{(L-1)}) = \sigma\big(\sum_s W_{st}^{(L)} h_s^{(L-1)}\big) \tag{5}$$

Note that in the computation of the conditional probability of the hidden unit $\boldsymbol{h}^{(l)}$, the probability incorporate both the lower hidden layer $\boldsymbol{h}^{(l-1)}$ and the upper hidden layer $\boldsymbol{h}^{(l+1)}$. Due to incorporate the more information from the lower and upper layers, the representational power of a DBM is more robust in the face of the noisy observed data (Karhunen et al., 2015). However, the character makes the conditional probability of DBM $P(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l-1)}, \boldsymbol{h}^{(l+1)})$ more complex than those of the DBN, $P(\boldsymbol{h}^{(l)}|\boldsymbol{h}^{(l+1)})$.

2.5. Generative Adversarial Networks

Due to their ability to learn deep representations without extensively annotated training data, Generative Adversarial Networks (GANs) have gained a lot of attention in computer vision and natural language processing (Goodfellow et al., 2014). GANs consist of two competing neural networks, a generator G and a discriminator D, as shown in **Figure 4**. The generator G parameterized by θ takes as input a random noise vector $\boldsymbol{z}$ from a prior distribution $p_{\boldsymbol{z}}(\boldsymbol{z};\theta)$ and outputs a sample $G(\boldsymbol{z})$, which can be regarded as a sample drawn from the generator data distribution p_g. The discriminator D that takes an input $G(\boldsymbol{z})$ or $\boldsymbol{x}$, and outputs the probability $D(\boldsymbol{x})$ or $D(\boldsymbol{G}(\boldsymbol{z}))$ to evaluate that the sample is from the generator G or the real data distribution. GANs simultaneously train the generator and discriminator where the generator G tries to generate realistic data to fool the discriminator, while the discriminator D tries to distinguish between the real and fake samples. Inspired by the game theory, the training process is to form a two-player minimax game with the value function $V(G, D)$ as follow:

$$\min_G \max_D V(G,D) = \mathbb{E}_{\boldsymbol{x}\sim p_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \tag{6}$$

where $p_{data}(\boldsymbol{x})$ denotes the real data distribution. After training alternately, if G and D have enough capacity, they will reach a point at which both cannot improve because $p_g = p_{data}$. In other words, the discriminator is unable to distinguish the difference between a real and a generated sample, i.e., $D(\boldsymbol{x}) = 0.5$. Although vanilla GAN has attracted considerable attention in various applications, there still remain several challenges related to training and evaluating GANs, such as model collapse and saddle points (Creswell et al., 2018). Therefore, many variants of GAN, such as Wasserstein GAN (WGAN) (Arjovsky et al., 2017)

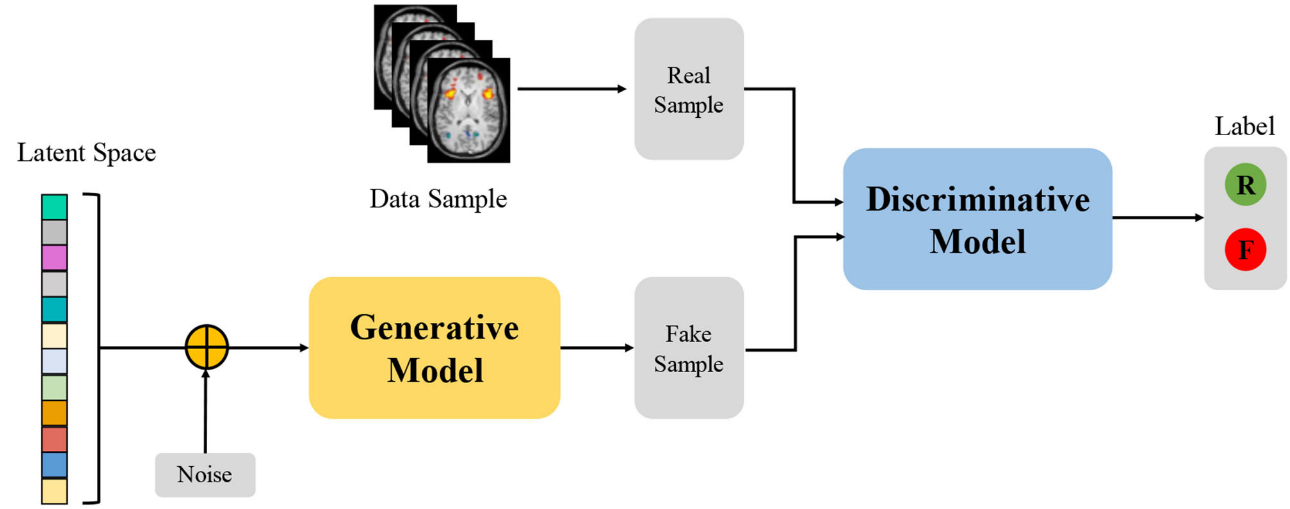


FIGURE 4 | Architecture of Generative Adversarial Networks. "R" and "F" represents the real and fake label, respectively.

and Deep Convolutional GAN (DCGAN) (Radford et al., 2015) have been proposed to overcome these challenges.

2.6. Convolutional Neural Networks

Compared to the SAE, DBN, and DBM, utilizing the inputs in vector form which inevitably destroys the structural information in images, the convolutional neural network (CNN) is designed to better retain and utilize the structural information among neighboring pixels or voxels and to required minimal preprocessing by directly taking two-dimensional (2D) or three-dimensional (3D) images as inputs (LeCun et al., 1998). Structurally, a CNN is a sequence of layers, and each layer of the CNN transforms one volume of activations to another through a differentiable function. **Figure 5** shows a typical CNN architecture (AlextNet model) for a computer vision task, which consists of three type neural layers: convolutional layers, pooling layers and fully connected layers (Krizhevsky et al., 2012). The convolutional layers are interspersed with pooling layers, eventually leading to the fully connected layers. The convolutional layer takes the pixels or voxels of a small patch of the input images, called the local receptive field and then utilizes various learnable kernels to convolve the receptive field to generate multiple feature maps. A pooling layer performs the non-linear downsampling to reduce the spatial dimensions of the input volume for the next convolutional layer. The fully connected layer input the 3D or 2D feature map to a 1D feature vector. The local response normalization is a non-trainable layer and performs a kind of "lateral inhibition" by normalizing over local input regions.

The major issue in training deep models is the over-fitting, which arises from the gap between the limited number of training samples and a large number of learnable parameters. Therefore, various techniques are designed to make the models train and generalize better, such as dropout and batch normalization to just name a few. A dropout layer randomly drops a fraction of the units or connections during each training iteration (Srivastava et al., 2014). It has also been demonstrated that dropout is able to successfully avoid over-fitting. In addition, batch normalization is another useful regularization and performs normalization with the running average of the mean–variance statistics of each mini-batch. It is shown that using batch normalization not only drastically speeds up the training time but also improves the generalization performance (Ioffe and Szegedy, 2015).

2.7. Graph Convolutional Networks

While the CNN has achieved huge success in extracting latent representations from Euclidean data (e.g., images, text, and video), there are a rapidly increasing number of various applications where data are generated from the non-Euclidean domain and needs to be efficiently analyzed. Researchers straightforwardly borrow ideas from CNN to design the architecture of graph convolutional networks (GCN) to handle complexity graph data (Kipf and Welling, 2016). **Figure 6** shows the process of a simple GCN with graph pooling layers for a graph classification task. The first step is to transform the traditional data to graph data, and the graph structure and node content information are therefore regarded as input. The graph convolutional layer plays a central role in extracting node hidden representations from aggregating the feature information from its neighbors. The graph pooling layers can be interleaved with the GCN layers and coarsened graphs into sub-graphs in order to obtained higher graph-level representations for each node on coarsened sub-graphs. After multiple fully connected layers, the softmax output layer is used to predict the class labels.

Depending on the types of graph convolutions, the GCN can be categorized into spectral-based and spatial-based methods. Spectral-based methods formulated graph convolution by introducing filters from the perspective of graph single processing. Spatial-based methods defined graph convolution directly on the graph, which operates on spatial close neighbors

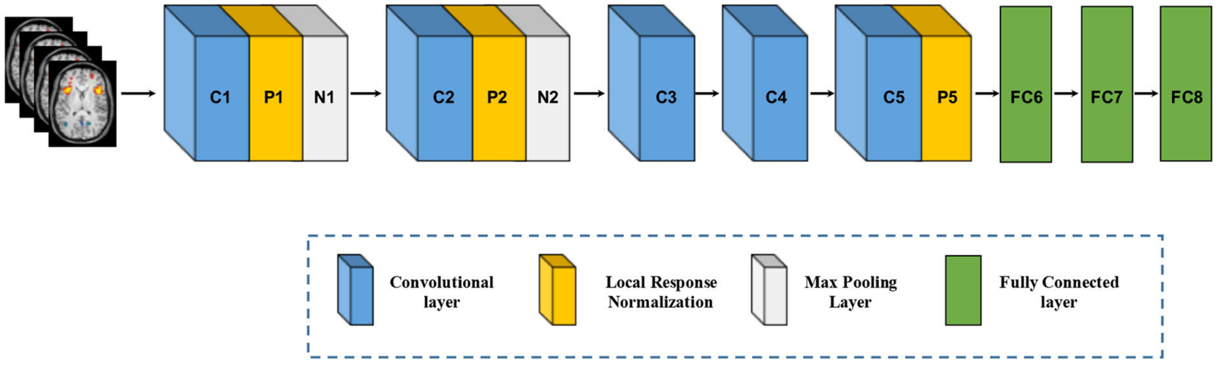


FIGURE 5 | Architecture of convolutional neural networks. Note that an implicit rectified linear unit (ReLU) non-linearity is applied after every layer. The natural images as input data in Krizhevsky et al. (2012) are replaced by brain MR images.

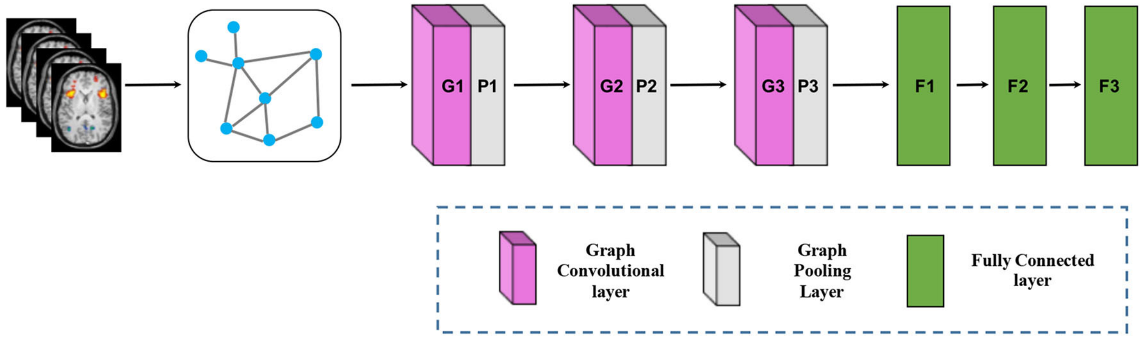


FIGURE 6 | Architecture of graph convolutional networks. To keep the figure simple, the softmax output layer is not shown.

to aggregate feature information. Due to drawbacks to spectral-based methods from three aspects, efficiency, generality, and flexibility, spatial-based methods have attracted more attention recently (Wu et al., 2019).

2.8. Recurrent Neural Networks

A recurrent neural network (RNN) is an extension of an FFNN, which is able to learn features and long-term dependencies from sequential and time-series data. The most popular RNN architecture is the long-short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), which is composed of a memory cell C_t, a forget gate f_t, an input gate i_t, and an output gate o_t (**Figure 7A**). The memory cell transfers relevant information all the way to the sequence chain, and these gates control the activation signals from various sources to decide which information is added to and removed from the memory cell. Unlike a basic RNN, the LSTM is able to decide whether to preserve the existing memory by the above-introduced gates. Theoretically, if the LSTM learns an important feature from the input sequential data, it can keep this feature over a long time, thus captures potential long-time dependencies. One popular LSTM variant is the Gated Recurrent Unit (GRU) (**Figure 7B**), which merges the forget and input gates into a single "update gate," and combines the memory cell state and hidden state into one state. The update gate decides how much information to add and throw away, and the reset gate decides how much previous information to forget. This makes the GRU is simpler than the standard LSTM (Cho et al., 2014).

2.9. Open Source Deep Learning Library

With the great successes of deep learning techniques in various applications, some famous research groups and companies have released their source codes and tools in deep learning. Due to these open source toolkits, people are able to easily build deep models for their applications even if they are not acquainted with deep learning techniques. **Supplementary Table 2** lists the most popular toolkits for deep learning and shows their main features.

3. APPLICATIONS IN BRAIN DISORDER ANALYSIS WITH MEDICAL IMAGES

The human brain is susceptible to many different disorders that strike at every stage of life. Developmental disorders usually first appear in early childhood, such as autism spectrum disorder and dyslexia. Although psychiatric disorders are typically diagnosed in teens or early adulthood, their origins may exist much earlier in life, such as depression and schizophrenia. Then, as people age, people become increasingly susceptible to Alzheimer's disease, Parkinson's disease, and other dementia diseases. In this section,

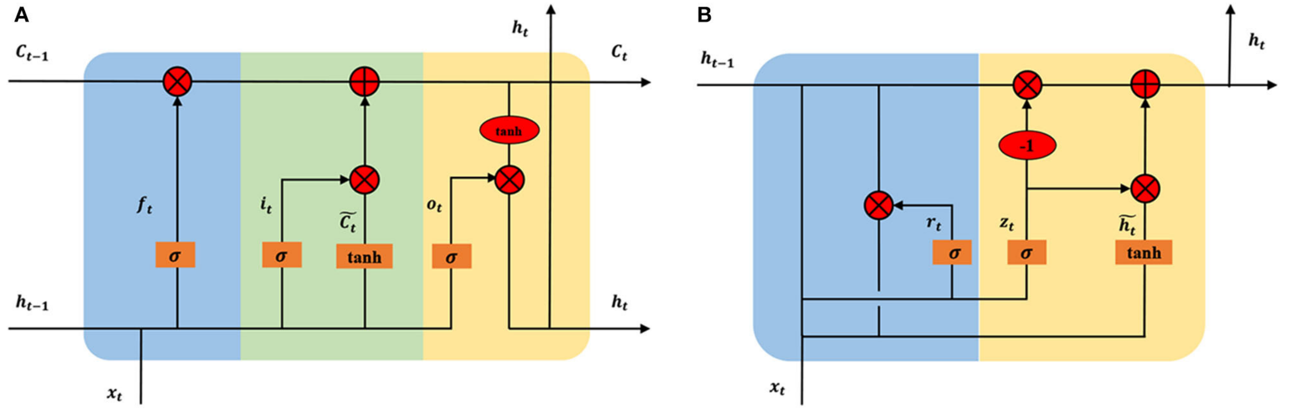


FIGURE 7 | Architectures of long short-term memory **(A)** and gated recurrent unit **(B)** In the subfigure **(A)**, the blue, green, and yellow represent the forget gate f_t, input, gate i_t, and output gate o_t, respectively. In the subfigure **(B)**, the blue and yellow represent the reset gate r_t and update gate z_t, respectively. x_t is input vector and h_t is the hidden state. To keep the figure simple, biases are not shown.

we select four typical brain disorders, including Alzheimer's disease, Parkinson's disease, Autism spectrum disorder and Schizophrenia. Alzheimer's disease and Parkinson's disease are both neurodegenerative disorders. Autism spectrum disorder and Schizophrenia are neurodevelopmental and psychiatric disorders, respectively.

3.1. Deep Learning for Alzheimer's Disease Analysis

Alzheimer's disease (AD) is a neurological, irreversible, progressive brain disorder and is the most common cause of dementia. Until now, the causes of AD are not yet fully understood, but accurate diagnosis of AD plays a significant role in patient care, especially at the early stage. For the study of AD diagnosis, the best-known public neuroimaging dataset is from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is a multi-site study that aims to improve clinical trials for the prevention and treatment of AD. The ADNI study has been running since 2004 and is now in its third phase (Mueller et al., 2005). Researchers collect, validate, and utilize data, including MRI and PET images, genetics, cognitive tests, cerebrospinal fluid (CSF), and blood biomarkers as predictors of the disease. Up to now, the ADNI dataset consists of ADNI-1, ADNI-GO, ADNI-2, and ADNI-3 and contains more than 1,000 patients. According to the Mini-Mental State Examination (MMSE) scores, these patients were in three stages of disease: normal control (NC), mild cognitive impairment (MCI), and AD. The MCI subject can be divided into two subcategories: converted MCI (cMCI) and stable MCI (sMCI), based on whether a subject converted to AD within a period of time (e.g., 24 months). The ADNI-GO and ADNI-2 provided two different MCI groups: early mild cognitive impairment (EMCI) and late mild cognitive impairment (LMCI), determined by a Wechsler Memory Scale (WMS) neuropsychological test.

Recently, plenty of papers have been published on the deep learning techniques for AD diagnosis. According to different architectures, these methods can be roughly divided into two subcategories: DGM-based and CNN-based methods. The DGM-based methods contained the DBN, DNM, SAE, and AE variants. Li et al. (2015) stacked multiple RBMs to construct a robust deep learning framework, which incorporated the stability selection and the multi-task learning strategy. Suk et al. (2014) proposed a series of methods based on deep learning models, such as the DBM and SAE (Suk et al., 2015, 2016). For example, the literature (Suk et al., 2015) applied the SAE to learn the latent representations from sMRI, PET, and CSF, respectively. Then, a multi-kernel SVM classifier was used to fuse the selected multi-modal features. Liu et al. (2015) also used SAE to extract features from multi-modal data, and a zero-masking strategy was then applied to fuse these learned features. Shi et al. (2017a) adopted multi-modality stacked denoising sparse AE (SDAE) to fuse cross-sectional and longitudinal features estimated from MR brain images. Lu et al. (2018) developed a multiscale deep learning network, which took the multiscale patch-wise metabolism features as input. This study was perhaps also the first study to utilize such a large number of FDG-PET images data. Martinez-Murcia et al. (2019) used a deep convolution AE (DCAE) architecture to extract features, which showed large correlations with clinical variables, such as age, tau protein deposits, and especially neuropsychological examinations. Due to small labeled samples in neuroimaging dataset, Shi et al. (2017b) proposed a multimode-stacked deep polynomial network (DPN) to effectively fuse and learn feature representation from a small multimodel neuroimaging data.

CNN-based methods learned all levels of features from raw pixels and avoided the manual ROIs annotation procedure and can be further subdivided into two subcategories: 2D-CNN and 3D-CNN. Gupta et al. (2013) pre-trained a 2D-CNN based on sMRI data through a sparse AE on random patches of natural images. The key technique was the use of cross-domain features to present MRI data. Liu and Shen (2014) used a similar strategy and pre-trained a pre-trained deep CNN on ImageNet. Sarraf et al. (2016) first used the fMRI data in deep learning applications. The 4D rs-fMRI and 3D MRI data were decomposed into 2D format images in the preprocessing step, and then the CNN-based architecture received these images in its input layer. Billones et al. designed a DemNet model based on the 16-layer VGGNet. The DemNet only selected the coronal image slices with indices 111–130 in 2D format images under the assumption that these slices covered the areas, which had the important features for the classification task (Billones et al., 2016). Liu et al. (2018b) proposed a novel classification framework that learned features from a sequence of 2D slices by decomposing 3D PET images. Then hierarchical 2D-CNN was built to capture the intra-slice features, while GRU was adopted to extract the inter-slice features.

The 3D brain images need to be decomposed into 2D slices in the preprocessing step, and this results in 2D-CNN methods discarding the spatial information. Many 3D-CNN methods were therefore proposed, and these can directly input 3D brain images. Payan and Montana (2015) pre-trained a 3D-CNN through a sparse AE on small 3D patches from sMRI scans. Hosseini-Asl et al. (2016) proposed a deep 3D-CNN, which was built upon a 3D CAE (Convolutional AE) to capture anatomical shape variations in sMRI scans. Liu et al. used multiple deep 3D-CNN on different local image patches to learn the discriminative features of MRI and PET images. Then, a set of upper high-level CNN was cascaded to ensemble the learned local features and discovered the latent multi-modal features for AD classification (Liu et al., 2018a). Karasawa et al. (2018) proposed deeper 3D-CNN architecture with 39 layers based on a residual learning framework (ResNet) to improve performance. Liu et al. (2018d) designed a landmark-based deep feature learning framework to learn the patch-level features, which were an intermediate scale between voxel-level and ROI-level. The authors firstly used a data-driven manner to identify discriminative anatomical landmarks from MR images, and they then proposed a 3D-CNN to learn patch-based features. This strategy can avoid the high-dimensional problem of voxel-level and manual definition of ROI-level. Subsequently, Liu et al. (2018c) developed a deep multi-instance CNN framework, where multiple image patches were used as a bag of instances to represent each specific subject, and then the label of each bag was given by the whole-image-level class label. To overcome the missing modality in multi-modal image data, Li et al. (2014) proposed a simple 3D-CNN to predict the missing PET images from the sMRI data. Results showed that the predicted PET data achieved similar classification accuracy to the true PET data. Additionally, the synthetic PET data and the real sMRI data obviously outperformed the single sMRI data. Pan et al. (2018) used Cycle-GAN to learn bi-directional mapping sMRI and PET to synthesize missing PET scans based on its corresponding sMRI scans. Then, landmark-based 3D-CNN was adapted for AD classification on the mixed image data. **Tables 1**, **2** summarized the statistic information of each paper reviewed above for AD diagnosis.

As an early stage of AD, MCI had a conversion rate as high as 10–15% per year in 5 years, but MCI was also the best time for treatment. Therefore, an effective predictive model construction for the early diagnosis of MCI had become a hot topic. Recently, some research based on GCN has been done for MCI prediction. Yu et al. (2019) and Zhao et al. (2019) both used the GCN, which combines neuroimaging information and the demographic relationship for MCI prediction. Song et al. (2019) implemented a multi-class the GCN classifier for classification of subjects on the AD spectrum into four classes. Guo et al. (2019) proposed PETNET model based on the GCN to analyzes PET signals defined on a group-wise inferred graph structure. **Tables 3**, **4** summarized the four papers for MCI prediction.

3.2. Deep Learning for Parkinson's Disease Analysis

Parkinson's disease (PD) is the most common neurodegenerative disorder after Alzheimer's disease, and it is provoked by progressive impairment and deterioration of neurons, caused by a gradually halt in the production of a chemical messenger in the brain. Parkinson's Progression Markers Initiative (PPMI) is an observational clinical study to verify progression markers in Parkinson's disease. The PPMI cohort comprises 400 newly diagnosed PD cases, 200 healthy, and 70 individuals that, while clinically diagnosed as PD cases, fail to show evidence of dopaminergic deficit. This latter group of patients is referred to as SWEDDs (Scans without Evidence of Dopamine Deficit) (Marek et al., 2011).

Some efforts based on deep learning have been done to design algorithms to help PD diagnosis. The Martinez-Murci team has continuously published a series of papers using deep learning techniques for PD diagnosis in a SPECT image dataset. Ortiz et al. (2016) designed a framework to automatically diagnose PD using deep sparse filtering-based features. Sparse filtering, based on ℓ_2-norm regularization, extracted the suitable features that can be used as the weight of hidden layers in a three-layer DNN. Subsequently, this team firstly applied 3D-CNN in PD diagnosis. These methods achieved up to a 95.5% accuracy and 96.2% sensitively (Martinez-Murcia et al., 2017). However, this 3D-CNN architecture with only two convolutional layers was too shallow and limited the capability to extract more discriminative features. Martinez-Murcia et al. (2018) therefore proposed a deep convolutional AE (DCAE) architecture for feature extraction. The DCAE overcome two common problems: the need for spatial normalization and the effect of imbalanced datasets. For a strongly imbalanced (5.69/1) PD dataset, DCAE achieved more than 93% accuracy. Choi et al. (2017) developed a deep CNN model (PDNet) consisted of four 3D convolutional layers. PDNet obtained high classification accuracy compared to the quantitative results of expert assessment and can further classify the SWEDD and NC subjects. Esmaeilzadeh et al. (2018) both utilized the sMRI scans and demographic information (i.e., age and gender) of patients to train a 3D-CNN model. The proposed method firstly found that the *Superior Parietal* part on the right hemisphere of the brain was critical in PD diagnosis. Sivaranjini and Sujatha (2019) directly introduced the AlexNet model, which was trained by the transfer learned network. Shen et al. (2019b) proposed an improved DBN model with an overlapping group lasso sparse penalty to learn useful low-level feature representations. To incorporate multiple brain neuroimaging modalities, Zhang et al. (2018b) and McDaniel

TABLE 1 | Overview of papers using deep learning techniques for AD diagnosis.

References	Year	Database	Subjects				Modality	Model
			AD	cMCI	sMCI	NC		
Suk et al. (2014)	2014	ADNI	93	76	128	101	sMRI + PET	DBM
Li et al. (2015)	2015	ADNI	51	43	56	52	sMRI + PET + CSF	DBN
Liu et al. (2015)	2015	ADNI	85	67	102	77	sMRI + PET	SAE
Suk et al. (2015)	2015	ADNI	51	43	56	52	sMRI + PET + CSF	SAE
Suk et al. (2016)	2016	ADNI	51	43	56	52	sMRI + PET + CSF	SAE
		–	198	167	236	229		
Shi et al. (2017a)	2017	ADNI	95	121		123	sMRI + Age	SDAE
Shi et al. (2017b)	2017	ADNI	51	43	56	52	sMRI + PET	DPN
Lu et al. (2018)	2018	ADNI	226	112	409	304	PET	SAE
Martinez-Murcia et al. (2019)	2019	ADNI	99	212		168	rs-fMRI	DCAE
Gupta et al. (2013)	2013	ADNI	200	411		232	sMRI	2D-CNN
Liu and Shen (2014)	2014	ADNI	200	411		232	sMRI	2D-CNN
Billones et al. (2016)	2016	ADNI	300	300		300	rs-fMRI	2D-CNN
Sarraf et al. (2016)	2016	ADNI	211	–	–	91	sMRI	2D-CNN
			52	–	–	92	rs-fMRI	
Liu et al. (2018b)	2017	ADNI	93	146		100	PET	2D-CNN + RNN
Payan and Montana (2015)	2015	ADNI	755	755		755	sMRI	3D-CNN
Hosseini-Asl et al. (2016)	2016	ADNI	70	70		70	sMRI	3D-CNN
Karasawa et al. (2018)	2018	ADNI	348	450	358	574	sMRI	3D-CNN
Liu et al. (2018a)	2018	ADNI	93	76	128	100	sMRI + PET	3D-CNN
Li et al. (2014)	2014	ADNI	193	167	236	229	sMRI + PET	3D-CNN
Liu et al. (2018c)	2018	ADNI	358	205	465	429	sMRI	3D-CNN
Liu et al. (2018d)	2018	ADNI	358	–	–	429	sMRI	3D-CNN
Pan et al. (2018)	2018	ADNI	358	205	465	429	sMRI + PET	3D-CNN + GAN

and Quinn (2019) both used a GCN model and presented an end-to-end pipeline without extra parameters involved for view pooling and pairwise matching. Transcranial sonography (TCS) had recently attracted increasing attention, and Shen et al. (2019a) proposed an improved DPN algorithm that embedded the empirical kernel mapping the network pruning strategy and dropout approach for the purposes of feature representation and classification for TCS-based PD diagnosis. **Table 5** summarized each paper above reviewed for PD diagnosis.

Up to now, only some papers have applied deep learning for PD diagnosis based on neuroimaging, and most of them adopt the 3D-CNN model. The traditional machine learning was still a popular and important technology for PD diagnosis, such as sparse feature learning (Lei et al., 2018), unsupervised learning (Singh and Samavedham, 2015), semi-unsupervised learning (Adeli et al., 2018), multi-task learning (Emrani et al., 2017), and classifier design (Shi et al., 2018).

3.3. Deep Learning for Austism Spectrum Disorder Analysis

Autism spectrum disorder (ASD) is a common neurodevelopmental disorder, which has affected 62.2 million ASD cases in the world in 2015. The Autism Imaging Data Exchange (ABIDE) initiative had aggregated rs-fMRI brain scans, anatomical and phenotypic datasets, collected from laboratories around the world. The ABIDE initiative included two large scale collections: ABIDE I and ABIDE II, which were released in 2012 and 2016, respectively. The ABIDE I collection involved 17 international sites and consisted of 1,112 subjects comprised of 539 from autism patients and 573 from NC. To further enlarge the number of samples with better-characterized, the ABIDE II collection involved 19 international sites, and aggregated 1,114 subjects from 521 individuals with ASD and 593 NC subjects (Di et al., 2014).

Many methods have been proposed on the application of deep learning for ASD diagnosis. These methods can be divided into three categories: AE-based methods, convolutional-based methods, and RNN-based methods. AE-based methods used various AE variations or stacked multiple AE to reduce data dimension and discovery highly discriminative representations. Hazlett et al. implemented the basic SAE, which primarily used surface area information from brain MRI at 6- and 12-months-old infants to predict the 24-months diagnosis of autism in children at high familial risk for autism. The SAE contained three hidden layers to reduce 315 dimension measurements to only two features (Hazlett et al., 2017). Two papers both used a stacked multiple sparse AE (SSAE) to learn low dimensional high-quality representations of functional connectivity patterns (Guo et al., 2017; Kong et al., 2019). But the difference was that

TABLE 2 | The classification performance of papers for AD diagnosis.

References	Accuracy (%)					
	AD/NC	AD/MCI	MCI/NC	cMCI/sMCI	3-ways[a]	4-ways[b]
Suk et al. (2014).	95.35 ± 5.23	–	85.67 ± 5.22	75.92 ± 15.37	–	–
Li et al. (2015)	91.4 ± 1.8	70.1 ± 2.3	77.4 ± 1.7	57.4 ± 3.6		
Liu et al. (2015)	91.4 ± 5.56	–	82.10 ± 4.91	–	–	53.79
Suk et al. (2015)	98.8 ± 0.9	83.7 ± 1.5	90.7 ± 1.2	83.3 ± 2.1	–	–
Suk et al. (2016)	95.09 ± 2.28	–	80.11 ± 2.64	74.15 ± 3.35	62.93	53.72
	90.27	–	70.86	73.93	57.74	47.83
Shi et al. (2017a)	91.95 ± 1.00	–	83.72 ± 1.16	–	–	–
Shi et al. (2017b)	97.13 ± 4.44	–	87.24 ± 4.52	76.88 ± 4.38	–	57.0±3.65
Lu et al. (2018)	93.58 ± 5.2	–	–	81.55 ± 7.42	–	–
Martinez-Murcia et al. (2019)	84.3 ± 6	–	–	71.5 ± 9	–	–
Gupta et al. (2013)	94.74	88.10	86.35	–	85.0	–
Liu and Shen (2014)	97.18 ± 1.5	94.51 ± 1.43	93.21 ± 1.02	–	91.72 ± 1.8	–
Billones et al. (2016)	98.33	93.89	91.67	–	91.85	–
Sarraf et al. (2016)	98.84/99.90	–	–	–	–	–
Liu et al. (2018b)	91.92	–	78.9	–	–	–
Payan and Montana (2015)	95.39	86.84	92.11	–	89.47	–
Hosseini-Asl et al. (2016)	99.3 ± 1.6	100	94.2 ± 2.0	–	94.8 ± 2.6	–
Karasawa et al. (2018)	94.0	–	90.0	–	87.0	–
Liu et al. (2018a)	93.26	–	73.34	–	–	–
Li et al. (2014)	92.87 ± 2.07	–	76.21 ± 2.05	72.44 ± 2.41	–	–
Liu et al. (2018c)	91.09	–	–	76.90	–	–
Liu et al. (2018d)	90.56	–	–	–	–	–
Pan et al. (2018)	92.50	–	–	79.06	–	–

[a] *3-ways represents the comparison: AD vs. NC vs. MCI.*
[b] *4-ways represents the comparison: AD vs. NC vs. cMCI vs. sMCI.*

TABLE 3 | Overview of papers using deep learning techniques for MCI prediction.

References	Year	Database	Subjects				Modality	Model
			NC	EMCI	LMCI	AD		
Zhao et al. (2019)	2019	ADNI	67	77	40	–	rs-fMRI	GCN
Yu et al. (2019)	2019	ADNI	44	44	38	–	rs-fMRI	GCN
Song et al. (2019)	2019	ADNI	12	12	12	12	DTI	GCN
Guo et al. (2019)	2019	ADNI	100	96	137	–	PET	GCN

TABLE 4 | The classification performance of papers for MCI prediction.

References	Accuracy (%)					
	EMCI/NC	LMCI/NC	EMCI/LMIC	MCI/NC	3-ways[a]	4-ways[b]
Zhao et al. (2019)	78.4	84.3	85.6	–	–	–
Yu et al. (2019)	87.5	89.02	79.27	–	–	–
Song et al. (2019)	–	–	–	–	–	89.0 ± 6
Guo et al. (2019)	–	–	–	93.0[c]	77.0	–

[a] *3-ways represents the comparison: NC vs. EMCI vs. LMCI.*
[b] *4-ways represents the comparison: NC vs. EMCI vs. LMCI vs. AD.*
[c] *MCI = ECMI + LMCI.*

TABLE 5 | Overview of papers using deep learning techniques for PD diagnosis.

References	Year	Database	Modality	Method	Modality			Accuracy (%)	
					PD	NC	SWEED	PD/NC	SWEED/NC
Ortiz et al. (2016)	2016	PPMI	SPECT	DNN	–	–	–	95.0	–
Martinez-Murcia et al. (2017)	2017	PPMI	SPECT	3D-CNN	158	111	32	95.5 ± 4.4	82.0 ± 6.8
Choi et al. (2017)	2017	PPMI	SPECT	3D-CNN	431	193	77	96.0	76.5
		SNUH[a]	SPECT		72	10	–	98.8	–
Esmaeilzadeh et al. (2018)	2018	PPMI	sMRI + DI[e]	3D-CNN	452	204	–	1.0	–
Martinez-Murcia et al. (2018)	2018	PPMI	SPECT	DCAE	1,110	195	–	93.3 ± 1.6	–
Sivaranjini and Sujatha (2019)	2019	PPMI	SPECT	2D-CNN	100	82	–	88.9	–
Zhang et al. (2018b)	2018	PPMI	sMRI + DTI	GCNN	596	158	–	95.37 (AUC)	–
McDaniel and Quinn (2019)	2019	PPMI	sMRI + DTI	GCNN	117	30	–	92.14	–
Shen et al. (2019b)	2019	HSHU[b]	PET	DBN	100	200	–	90.0	–
		WXH[c]	PET		25	25	–	86.0	–
Shen et al. (2019a)	2019	Multi-site[d]	TCS	DPN	76	77	–	86.95 ± 3.15	–

[a]SNUH, Seoul National University Hospital cohort. [b]HSH, HuaShan Hospital cohort. [c]WXH, WuXi 904 Hospital cohort. [d]Shanghai East Hospital of Tongji University and the Second Affiliated Hospital of Soochow University. [e]DI, Demographic Information.

Guo et al. input the whole-brain functional connectivity patterns and Kong et al. only selected the top 3,000 ranked connectivity features by F-score in descending order. Dekhil et al. (2018) built an automated autism diagnosis system, which used 34 sparse AE for 34 spatial activation areas. Each sparse AE extracted the power spectral densities (PSDs) of time courses in a higher-level representation and simultaneously reduced the feature vectors dimensionality. Choi (2017) used VAE to summarize the functional connectivity networks into two-dimensional features. One feature was identified using a high discrimination between ASD and NC, and it was closely associated with ASD-related brain regions. Heinsfeld et al. (2018) used DAE to reduce the effect of multi-site heterogeneous data and improve the generalization. Due to insufficient training samples, Li et al. (2018a) developed a novel deep neural network framework with the transfer learning technique for enhancing ASD classification. This framework was firstly trained an SSAE to learn functional connectivity patterns from healthy subjects in the existing databases. The trained SSAE was then transferred to a new classification with limited target subjects. Saeed et al. designed a data augmentation strategy to produce synthetic datasets needed for training the ASD-DiagNet model. This model was composed of an AE and a single-layer perceptron to improve the quality of extracted features (Saeed et al., 2019).

Due to collapsed the rs-fMRI scans into a feature vector, the above methods discarded the spatial structure of the brain networks. To fully utilize the whole brain spatial fMRI information, Li et al. (2018b) implemented 3D-CNN to capture spatial structure information and used sliding windows over time to measure temporal statistics. This model was able to learn ASD-related biological markers from the output of the middle convolution layer. Khosla et al. proposed a 3D-CNN framework for connectome-based classification. The functional connectivity of each voxel to various target ROIs was used as input features, which reserved the spatial relationship between voxels. Then the ensemble learning strategy was employed to average the different ROI definitions to reduce the effect of empirical selections, it and obtained more robust and accurate results (Khosla et al., 2018). Ktena et al. (2018) implemented a Siamese GCN to learn a graph-similarity metric, which took the graph structure into consideration for the similarity between a pair of graphs. This was the first application of metric learning with graph convolutions on brain connectivity networks. Parisot et al. (2017) introduced a spectral GCN for brain analysis in populations combining imaging and non-imaging information. The populations were represented as a sparse graph where each vertex corresponded to an imaging feature vector of a subject, and the edge weights were associated with phenotypic data, such as age, gender, and acquisition sites. Like the graph-based label propagation, a GCN model was used to infer the classes of unlabeled nodes on the partially labeled graphs. There existed no definitive method to construct reliable graphs in practice. Thus, Anirudh and Thiagarajan (2017) proposed a bootstrapped version of GCN to reduce the sensitivity of models on the initial graph construction step. The bootstrapped GCN used an ensemble of the weekly GCN, each of which was trained by a random graph. In addition, Yao et al. (2019) proposed a multi-scale triplet GCN to avoid the spatial limitation of a single template. A multi-scale templates for coarse-to-fine ROI parcellation were applied to construct multi-scale functional connectivity patterns for each subject. Then a triple GCN model was developed to learn multi-scale graph features of brain networks.

Several RNN-based methods were proposed to fully utilize the temporal information in the rs-fMRI time-series data. Bi et al. (2018) designed a random NN cluster, which combined multiple NNs into a model, to improve the classification performance in the diagnosis of ASD. Compared to five different NNs, the random Elman cluster obtained the highest accuracy. It is because that the Elman NN fit handling the dynamic data. Dvornek et al. (2017) first applied LSTM to ASD classification,

TABLE 6 | Overview of papers using deep learning techniques for ASD diagnosis.

References	Year	Database	Subject		Modality	Model	Accuracy (%)
			ASD	NC			
Guo et al. (2017)	2017	ABIDE I	55	55	rs-fMRI	SSAE	86.36
Kong et al. (2019)	2019	ABIDE I	78	104	rs-fMRI	SSAE	90.39
Li et al. (2018a)	2018	ABIDE: UM[a]	48	65	rs-fMRI	SSAE	67.2
		ABIDE:UCLA[b]	36	39			62.3
		ABIDE: USM[c]	38	23			70.4
		ABIDE: LEUVEN[d]	27	34			68.3
Choi (2017)	2017	ABIDE	465	507	rs-fMRI	VAE	0.60 (AUC)
Heinsfeld et al. (2018)	2018	ABIDE	505	530	rs-fMRI	DAE	70.0
Hazlett et al. (2017)	2017	NDAR[e]	106	42	rs-fMRI	SAE	88.0
Dekhil et al. (2018)	2018	NDAR	123	160	rs-fMRI	SSAE	91.0 ± 3.2
Saeed et al. (2019)	2019	ABIDE	505	530	rs-fMRI	AE	70.1 ± 3.2
Li et al. (2018b)	2018	–	82	48	rs-fMRI	3D-CNN	89.0 ± 5.0 (F-score)
Khosla et al. (2018)	2018	ABIDE	542	625	rs-fMRI	3D-CNN	73.3
(Parisot et al., 2017)	2017	ABIDE	403	468	rs-fMRI	GCN	69.5
Anirudh and Thiagarajan (2017)	2017	ABIDE	404	468	rs-fMRI	GCN	70.8
Yao et al. (2019)	2019	ABIDE	438	544	rs-fMRI	GCN	67.3
Ktena et al. (2018)	2018	ABIDE	403	468	rs-fMRI	GCN	62.9
Dvornek et al. (2017)	2017	ABIDE	1,100	–	rs-fMRI	LSTM	68.5 ± 5.5
Bi et al. (2018)	2018	ABIDE	50	42	rs-fMRI	RNN	84.7 ± 3.2

[a]University of Michigan. [b]University of California, Los Angeles. [c]University of Utah School of Medicine. [d]Katholieke Universiteit Leuven. [e]National Database of Autism Research.

which directly used the rs-fMRI time-series data, rather than the pre-calculated measures of brain functional connectively. The authors thought that the rs-fMRI time-series data contained more useful information of dynamic brain activity than single and static functional connectivity measures. For clarity, the important information of the above-mentioned papers was summarized in **Table 6**.

3.4. Deep Learning for Schizophrenia Analysis

Schizophrenia (SZ) is a prevalent psychiatric disorder and affects 1% of the population worldwide. Due to the complex clinical symptoms, the pathological mechanism of schizophrenia remains unclear and there is no definitive standard in the diagnosis of SZ. Different from the ADNI for AD diagnosis, the PPMI for PD diagnosis, and the ABIDE for ASD diagnosis, there was not a widely used neuroimaging dataset for the SZ diagnosis. Therefore, some studies have successfully applied source datasets that were available from the medical research centers, universities, and hospitals.

Recently, some studies have successfully applied deep learning algorithms to SZ diagnosis and have seen significant improvement. These methods were divided into two categories: unimodality and multi-modality, according to the types of input data, rather than according to deep learning architectures like AD or ASD diagnosis.

The unimodality category only used a single type of MRI and can furthermore be classified into subclasses: sMRI-methods and fMRI-methods. sMRI-methods discovery latent features from sMRI dataset, which can provide information on the tissue structure of the brain, such as gray matter, white matter, and cerebrospinal fluid. Plis et al. and Pinaya et al. used the DBN model, which only contained three hidden layers, to automatically extract feature for SZ identification. The results achieved a modestly higher predictive performance than the shallow-architecture SVM approach (Plis et al., 2014; Pinaya et al., 2016). Different from the DBN model in Pinaya et al. (2016), Pinaya et al. (2019) trained an SAE to create a normative model from 1,113 NC subjects, then used this model to estimate total and regional neuroanatomical deviation in individual patients with SZ. Ulloa et al. proposed a novel classification architecture that used synthetic sMRI scans to mitigate the effects of a limited sample size. To generate synthetic samples, a data-driven simulator was designed that can capture statistical properties from observed data using independent component analysis (ICA) and a random variable sampling method. Then a 10-layer DNN was trained exclusively on continuously generated synthetic data, and it greatly improves generalization in the classification of SZ patients and NC (Ulloa et al., 2015).

The fMRI-methods extracted discriminative features from rs-fMRI brain images with functional connectivity networks. Kim et al. (2015) learned lower-to-higher features via the DNN model in which each hidden layer was added L_1-regularization to control the weight sparsity, and they also achieved 85.8% accuracy. Patel et al. used an SAE model with four hidden layers to separately train on each brain region. The input layer directly uses the complete time series of all active voxels without converting them into region-wise mean

TABLE 7 | Overview of papers using deep learning techniques for SZ diagnosis.

References	Year	Database	Subject		Modality	Model	Accuracy (%)
			SZ	NC			
Plis et al. (2014)	2014	Multi-site1[a]	198	191	sMRI	DBN	91.0 + 14 (F-score)
Ulloa et al. (2015)	2015	Multi-site1	198	191	sMRI	DNN	75.0 ± 4 (AUC)
Pinaya et al. (2016)	2016	UNIFESP[b]	143	83	sMRI	DBN	73.55 ± 6.84
Pinaya et al. (2019)	2019	NUSDAST[c]	30	40	sMRI	SAE	70.7
Kim et al. (2015)	2015	NITRC[d]	50	50	rs-fMRI	DNN	85.8
Patel et al. (2016)	2016	COBRE[e]	72	74	rs-fMRI	SAE	92.0
Zeng et al. (2018)	2018	Multi-site2[f]	357	377	rs-fMRI	SAE	85.0 ± 1.2
Qureshi et al. (2019)	2019	COBRE	72	74	rs-fMRI	3D-CNN	98.09 ± 1.01
Dakka et al. (2017)	2017	FBIRN[g]	46	49	rs-fMRI	CNN + LSTM	66.4
Yan et al. (2019)	2019	Multi-site3[h]	558	542	rs-fMRI	CNN + GRU	83.2 ± 3.2
Qi and Tejedor (2016)	2016	MLSP2014	69	75	sMRI + fMRI	DCCA/DCCAE	94.2/95.0 (AUC)
Srinivasagopalan et al. (2019)	2019	MLSP2014	69	75	sMRI + fMRI	DNN	94.44
Ulloa et al. (2018)	2018	FBIRN	135	169	sMRI + fMRI	DNN	85.0 ± 5.0 (AUC)

[a]Johns Hopkins University; the Maryland Psychiatric Research Center; the Institute of Psychiatry; the Western Psychiatric Institute and Clinic at the University of Pittsburgh.
[b]the Universidade Federal de São Paulo.
[c]Northwestern University Schizophrenia Data and Software Tool.
[d]Neuroimaging Informatics Tools and Resources Clearinghouse website.
[e]Center for Biomedical Research Excellence.
[f]Xijing Hospital; First Affliated Hospital of Anhui Medical University; Second Xiangya Hospital; COBRE; the University of California, Los Angles and Washington University School of Medicine.
[g]The Function Biomedical Informatics Research Network Data.
[h]Peking University Sixth Hospital; Beijing Huilongguan Hospital; Xinxiang Hospital; Xinxiang Hospital; Xijing Hospital; Renmin Hospital of Wuhan University; Zhumadian Psychiatric Hospital.

time series. This therefore ensured that the model retained more information (Patel et al., 2016). Due to the limited size of SZ dataset, Zeng et al. collected a large multi-site rs-fMRI dataset from seven neuroimaging resources. An SAE with an optimized discriminant item was designed to learn imaging site-shared functional connectivity features. This model can achieve accurate SZ classification performance across multiple independent imaging sites, and the learned features found that dysfunctional integration of the cortical-striatal-cerebellar circuit may play an important role in SZ (Zeng et al., 2018). Qureshi et al. built a 3D-CNN-based deep learning classification framework, which used the 3D ICA functional network maps as input. These ICA maps served as highly discriminative 3D imaging features for the discrimination of SZ (Qureshi et al., 2019). To exploit both spatial and temporal information, Dakka et al. and Yan et al. proposed a recurrent convolutional neural network involving CNN followed by LSTM and GRU, respectively. The CNN extracted spatial features, which then were fed to the followed RNN model to learn the temporal dependencies (Dakka et al., 2017; Yan et al., 2019).

Combined multi-modality brain images can improve the performance of disorder diagnosis. The MLSP2014 (Machine Learning for Signal Processing) SZ classification challenge provided 75 NC and 69 SZ, which both contained sMRI and rs-fMRI brain images. Qi and Tejedor (2016) used deep canonical correlation analysis (DCCA) and deep canonically correlated auto-encoders (DCCAE) to fuse multi-modality features. But in the proposed method, two modalities features directly were combined as 411 dimensional vector, then fed to the three-layer DNN model (Srinivasagopalan et al., 2019). To alleviate the missing modality, the synthetic sMRI and rs-fMRI images were generated by a generator proposed, and they were then used to train a multi-modality DNN (Ulloa et al., 2018). For clarity, the important information of the above-mentioned papers was summarized in **Table 7**. From this table, it can be seen the datasets for SZ diagnosis come from different universities, hospitals, and medical centers.

4. DISCUSSION AND FUTURE DIRECTION

As can be seen from this survey, consideration research has been reviewed on the subject of deep learning across four brain disorder diseases. Furthermore, the number of publications on medical imaging analysis shows an almost exponential growth in PubMed. Unfortunately, there is no unified deep learning framework that could be generally used for every disease research, even only for human disorder diseases. This is consistent with the "*No Free Lunch*" theorem, which states that there is no one model that works best for every problem. Thus, different deep learning methods are developed using different imaging modalities for a disease-specific task.

Although deep learning models have achieved great success in the field of neuroimaging-based brain disorder analysis, there are still some challenges that deserve further investigation. We summarize these potential challenges as follows and explore possible solutions.

First, deep learning algorithms highly depend on the configuration of hyper-parameter, which may dramatically fluctuate the performance. The hyper-parameter set composed of two parts: model optimization parameters (e.g., the optimization method, learning rate, and batch sizes, etc.) and network structure parameters (e.g., number of hidden layers and units, dropout rate, activation function, etc.). To obtain the best configuration, hyper-parameter optimization methods, including manual (e.g., grid search and random search) and automatic (e.g., Bayesian Optimization), are proposed. However, the method behind designing the architecture of deep neural networks still depends on the experienced experts. Recently, neural architecture search (NAS) automates this design of network architecture and indeed received new state-of-the-art performance (Zoph and Le, 2016; He et al., 2019). Additionally, another interesting technique called Population-Based Training (PTB), which is inspired by genetic algorithms, bridges and extends parallel search methods and sequential optimization methods. PBT is ability to automatic discovery of hyper-parameter schedules and model selection, which leads to stable training and better final performance (Jaderberg et al., 2017). It indicates that the hyper-parameter optimization may further mine the potential of deep learning in medical analysis.

Second, deep neural networks rely on complicated architectures to learn feature representations of the training data, and then makes its predictions for various tasks. These methods can achieve extremely accurate performances and may even beat human experts. But it is difficult to trust these predictions based on features you cannot understand. Thus, the black-box natural of the deep learning algorithms has restricted the practical clinical use. Some studies begin to explore the interpretability of deep learning in medical image analysis, and aim to show the features that most influence the predictions (Singh et al., 2020). An attention-based deep learning method is proposed and deemed as an interpretable tool for medical image analysis, which inspired by the way human pay attention to different parts of an image or the disease's influence on different regions of neuroimages (Sun et al., 2019b; Huang et al., 2020). The clinical diagnosis information as a modality is fused into the model to improve accuracy as well as give more comprehensive interpretability of outcomes (Hao et al., 2016, 2017; Wang et al., 2019a). Thus, how to improve the interpretability of deep learning model is worth further study and attention.

Third, deep learning methods require a large number of samples to train neural networks, though it is usually difficult to acquire training samples in many real-world scenarios, especially for neuroimaging data. The lack of sufficient training data in neuroimage analysis has been repeatedly mentioned as a challenge to apply deep learning algorithms. To address this challenge, a data augmentation strategy has been proposed, and it is widely used to enlarge the number of training samples (Hussain et al., 2017; Shorten and Khoshgoftaar, 2019). In addition, the use of transfer learning (Cheng et al., 2015, 2017) provides another solution by transferring well-trained networks on big sample datasets (related to the to-be-analyzed disease) to a small sample dataset for further training.

Fourth, the missing data problem is unavoidable in multimodal neuroimaging studies, because subjects may lack some modalities due to patient dropouts and poor data quality. Conventional methods typically discard data-missing subjects, which will significantly reduce the number of training subjects and degrade the diagnosis performance. Although many data-imputing methods have been proposed, most of them focus on imputing missing hand-crafted feature values that are defined by experts for representing neuroimages, while the hand-crafted features themselves could be not discriminative for disease diagnosis and prognosis. Several recent studies (Pan et al., 2018, 2019) propose that we directly impute missing neuroimages (e.g., PET) based on another modality neuroimages (e.g., MRI), while the correspondence between imaging data and non-imaging data has not been explored. We expect to see more deep network architectures in the near future to explore the association between different data modalities for imputing those missing data.

Fifth, an effective fusion of multimodal data has always been a challenge in the field. Multimodal data reflects the morphology, structure, and physiological functions of normal tissues and organs from different aspects and has strong complementary characteristics between different models. Previous studies for multimodal data fusion can be divided into two categories, *data-level fusion* (focus on how to combine data from different modalities) and *decision-level fusion* (focus on ensembling classifiers). Deep neural network architectures allow a third form of multimodal fusion, i.e., the intermediate fusion of learned representations, offering a truly flexible approach to multimodal fusion (Hao et al., 2020). As deep-learning architectures learn a hierarchical representation of underlying data across its hidden layers, learned representations between different modalities can be fused at various levels of abstraction. Further investigation is desired to study which layer of deep integration is optimal for problems at hand.

Furthermore, different imaging modalities usually reflect different temporal and spatial scales information of the brain. For example, sMRI data reflect minute-scale time scales information of the brain, while fMRI data can provide second-scale time scales information. In the practical diagnosis of brain disorder, it shows great significance for the implementation of early diagnosis and medical intervention by correctly introducing the spatial relationship of the diseased brain regions and other regions and the time relationship of the development of the disease progress (Jie et al., 2018; Zhang et al., 2018a). Although previous studies have begun to study the pathological mechanisms of brain diseases on a broad temporal and spatial scales, those methods usually consider either temporal or spatial characteristics (Wang et al., 2019b,d). It is therefore desirable to develop a series of deep learning frameworks to fuse temporal and spatial information for automated diagnosis of brain disorder.

Finally, the utilization of multi-site data for disease analysis has recently attracted increased attention (Heinsfeld et al., 2018;

Wang et al., 2018, 2019c) since a large number of subjects from multiple imaging sites are beneficial for investigating the pathological changes of disease-affected brains. Previous methods often suffer from inter-site heterogeneity caused by different scanning parameters and subject populations in different imaging sites by assuming that these multi-site data are drawn from the same data distribution. Constructing accurate and robust learning models using heterogeneous multi-site data is still a challenging task. To alleviate the inter-site data heterogeneity, it could be a promising way to simultaneously learn adaptive classifiers and transferable features across multiple sites.

5. CONCLUSION

In this paper, we reviewed the most recent studies on the subject of applying the deep learning techniques in neuroimaging-based brain disorder analysis and focused on four typical disorders. AD and PD are both neurodegenerative disorders. ASD and SZ are neurodevelopmental and psychiatric disorders, respectively. Deep learning models have achieved state-of-the-art performance across the four brain disorders using brain images. Finally, we summarize these potential challenges and discuss possible research directions. With the clearer pathogenesis of human brain disorders, the further development of deep learning techniques, and the larger size of open-source datasets, a human-machine collaboration for medical diagnosis and treatment will ultimately become a symbiosis in the future.

AUTHOR CONTRIBUTIONS

DZ, ML, and LZ designed this review. LZ and MW searched the literatures. LZ wrote this manuscript. All authors read, edited, and discussed the article.

FUNDING

This work was supported in part by National Natural Science Foundation of China (NSFC) under grant (Nos. 61802193, 61861130366, 61876082, and 61732006), the National Key R&D Program of China (Grant Nos. 2018YFC2001600, 2018YFC2001602, and 2018ZX10201002), the Natural Science Foundation of Jiangsu Province under grant (No. BK20170934), the Royal Society-Academy of Medical Sciences Newton Advanced Fellowship (No. NAF\R1\180371), and the Fundamental Research Funds for the Central Universities (No. NP2018104).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00779/full#supplementary-material

REFERENCES

Adeli, E., Thung, K.-H., An, L., Wu, G., Shi, F., Wang, T., et al. (2018). Semi-supervised discriminative classification robust to sample-outliers and feature-noises. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 515–522. doi: 10.1109/TPAMI.2018.2794470

Anirudh, R., and Thiagarajan, J. J. (2017). Bootstrapping graph convolutional neural networks for autism spectrum disorder classification. *arXiv* 1704.07487.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv* 1701.07875.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Shanghai: IEEE), 4945–4949. doi: 10.1109/ICASSP.2016.7472618

Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H., and Montreal, U. (2007). "Greedy Layer-Wise Training of Deep Networks," in *Advances in Neural Information Processing Systems* (Vancouver, BC: ACM), 153–160. doi: 10.5555/2976456.2976476

Bi, X., Liu, Y., Jiang, Q., Shu, Q., Sun, Q., and Dai, J. (2018). The diagnosis of autism spectrum disorder based on the random neural network cluster. *Front. Hum. Neurosci.* 12:257. doi: 10.3389/fnhum.2018.00257

Billones, C. D., Demetria, O. J. L. D., Hostallero, D. E. D., and Naval, P. C. (2016). "DemNet: a convolutional neural network for the detection of Alzheimer's disease and mild cognitive impairment," in *Region 10 Conference, 2016 IEEE* (Singapore: IEEE), 3724–3727. doi: 10.1109/TENCON.2016.7848755

Brody, H. (2013). Medical imaging. *Nature* 502:S81. doi: 10.1038/502S81a

Cheng, B., Liu, M., Shen, D., Li, Z., and Zhang, D. (2017). Multi-domain transfer learning for early diagnosis of Alzheimer's disease. *Neuroinformatics* 15, 115–132. doi: 10.1007/s12021-016-9318-5

Cheng, B., Liu, M., Suk, H.-I., Shen, D., and Zhang, D. (2015). Multimodal manifold-regularized transfer learning for MCI conversion prediction. *Brain Imaging Behav.* 9, 913–926. doi: 10.1007/s11682-015-9356-x

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: encoder-decoder approaches. *arXiv* 1409.1259. doi: 10.3115/v1/W14-4012

Choi, H. (2017). Functional connectivity patterns of autism spectrum disorder identified by deep feature learning. *arXiv* 1707.07932.

Choi, H., Ha, S., Im, H. J., Paek, S. H., and Lee, D. S. (2017). Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. *Neuroimage Clin.* 16, 586–594. doi: 10.1016/j.nicl.2017.09.010

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: an overview. *IEEE Signal Process. Mag.* 35, 53–65. doi: 10.1109/MSP.2017.2765202

Dakka, J., Bashivan, P., Gheiratmand, M., Rish, I., Jha, S., and Greiner, R. (2017). Learning neural markers of schizophrenia disorder using recurrent neural networks. *arXiv* 1712.00512.

Dekhil, O., Hajjdiab, H., Shalaby, A., Ali, M. T., Ayinde, B., Switala, A., et al. (2018). Using resting state functional MRI to build a personalized autism diagnosis system. *PLoS ONE* 13:e0206351. doi: 10.1371/journal.pone.0206351

Di, M. A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19, 659–667. doi: 10.1038/mp.2013.78

Durstewitz, D., Koppe, G., and Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Mol. Psychiatry* 24, 1583–1598. doi: 10.1038/s41380-019-0365-9

Dvornek, N. C., Ventola, P., Pelphrey, K. A., and Duncan, J. S. (2017). "Identifying autism from resting-state fMRI using long short-term memory networks," in *International Workshop on Machine Learning in Medical Imaging* (Quebec City, QC: Springer), 362–370. doi: 10.1007/978-3-319-67389-9_42

Emrani, S., McGuirk, A., and Xiao, W. (2017). "Prognosis and diagnosis of Parkinson's disease using multi-task learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS: ACM), 1457–1466. doi: 10.1145/3097983.3098065

Erickson, B. J., Korfiatis, P., Akkus, Z., and Kline, T. L. (2017). Machine learning for medical imaging. *Radiographics* 37, 505–515. doi: 10.1148/rg.2017160130

Esmaeilzadeh, S., Yang, Y., and Adeli, E. (2018). End-to-end Parkinson disease diagnosis using brain MR-images by 3D-CNN. *arXiv* 1806.05233.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning, Vol. 1*. 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*, 2672–2680. Available online at: https://papers.nips.cc/paper/5423-generative-adversarial-nets

Guo, J., Qiu, W., Li, X., Zhao, X., Guo, N., and Li, Q. (2019). Predicting Alzheimer's disease by hierarchical graph convolution from positron emission tomography imaging. *arXiv* 1910.00185. doi: 10.1109/BigData47090.2019.9005971

Guo, X., Dominick, K. C., Minai, A. A., Li, H., Erickson, C. A., and Lu, L. J. (2017). Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. *Front. Neurosci*. 11:460. doi: 10.3389/fnins.2017.00460

Gupta, A., Ayhan, M., and Maida, A. (2013). "Natural image bases to represent neuroimaging data," in *International Conference on Machine Learning* (Atlanta, GA), 987–994.

Hao, X., Bao, Y., Guo, Y., Yu, M., Zhang, D., Risacher, S. L., et al. (2020). Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease. *Med. Image Anal*. 60:101625. doi: 10.1016/j.media.2019.101625

Hao, X., Li, C., Du, L., Yao, X., Yan, J., Risacher, S. L., et al. (2017). Mining outcome-relevant brain imaging genetic associations via three-way sparse canonical correlation analysis in Alzheimer's disease. *Sci. Rep*. 7:44272. doi: 10.1038/srep44272

Hao, X., Yao, X., Yan, J., Risacher, S. L., Saykin, A. J., Zhang, D., et al. (2016). Identifying multimodal intermediate phenotypes between genetic risk factors and disease status in Alzheimer's disease. *Neuroinformatics* 14, 439–452. doi: 10.1007/s12021-016-9307-8

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/nature21369

He, X., Zhao, K., and Chu, X. (2019). AutoML: a survey of the state-of-the-art. *arXiv* 1908.00709.

Heidenreich, A., Desgrandschamps, F., and Terrier, F. (2002). Modern approach of diagnosis and management of acute flank pain: review of all imaging modalities. *Eur. Urol*. 41, 351–362. doi: 10.1016/S0302-2838(02)00064-7

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuroimage Clin*. 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Hinton, G., and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science* 268, 1158–1161. doi: 10.1126/science.7761831

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

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput*. 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Hosseini-Asl, E., Gimel'farb, G., and El-Baz, A. (2016). Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network. *arXiv* 1607.00556.

Hu, Z., Tang, J., Wang, Z., Zhang, K., Zhang, L., and Sun, Q. (2018). Deep learning for image-based cancer detection and diagnosis'a survey. *Pattern Recogn*. 23, 134–149. doi: 10.1016/j.patcog.2018.05.014

Huang, J., Zhou, L., Wang, L., and Zhang, D. (2020). Attention-diffusion-bilinear neural network for brain network analysis. *IEEE Trans. Med. Imaging* 39, 2541–2552. doi: 10.1109/TMI.2020.2973650

Hussain, Z., Gimenez, F., Yi, D., and Rubin, D. (2017). "Differential data augmentation techniques for medical imaging classification tasks," in *AMIA Annual Symposium Proceedings. AMIA Symposium 2017* (Washington, DC), 979–984.

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

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., et al. (2017). Population based training of neural networks. *arXiv* 1711.09846.

Jie, B., Liu, M., Lian, C., Shi, F., and Shen, D. (2018). "Developing novel weighted correlation kernels for convolutional neural networks to extract hierarchical functional connectivities from fMRI for disease diagnosis," in *International Workshop on Machine Learning in Medical Imaging* (Granada: Springer), 1–9. doi: 10.1007/978-3-030-00919-9_1

Karasawa, H., Liu, C.-L., and Ohwada, H. (2018). "Deep 3D convolutional neural network architectures for Alzheimer's disease diagnosis," in *Asian Conference on Intelligent Information and Database Systems* (Dong Hoi City: Springer), 287–296. doi: 10.1007/978-3-319-75417-8_27

Karhunen, J., Raiko, T., and Cho, K. (2015). "Unsupervised deep learning: a short review," in *Advances in Independent Component Analysis and Learning Machines* (Elsevier), 125–142. doi: 10.1016/B978-0-12-802806-3.00007-5

Khosla, M., Jamison, K., Kuceyeski, A., and Sabuncu, M. R. (2018). "3D convolutional neural networks for classification of functional connectomes," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (Granada: Springer), 137–145. doi: 10.1007/978-3-030-00889-5_16

Kim, J., Calhoun, V. D., Shim, E., and Lee, J. H. (2015). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *Neuroimage* 124, 127–146. doi: 10.1016/j.neuroimage.2015.05.018

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

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv* 1609.02907.

Kong, Y., Gao, J., Xu, Y., Pan, Y., Wang, J., and Liu, J. (2019). Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. *Neurocomputing* 324, 63–68. doi: 10.1016/j.neucom.2018.04.080

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, eds M. I. Jordan, Y. LeCun, and S. A. Solla (Lake Tahoe, NV: ACM), 1097–1105. doi: 10.1145/3065386

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., et al. (2018). Metric learning with spectral graph convolutions on brain connectivity networks. *Neuroimage* 169, 431–442. doi: 10.1016/j.neuroimage.2017.12.052

Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. (2009). Exploring strategies for training deep neural networks. *J. Mach. Learn. Res*. 10, 1–40. doi: 10.1145/1577069.1577070

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

Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., et al. (2017). Deep learning in medical imaging: general overview. *Korean J. Radiol*. 18, 570–584. doi: 10.3348/kjr.2017.18.4.570

Lei, H., Zhao, Y., Wen, Y., Luo, Q., Cai, Y., Liu, G., et al. (2018). Sparse feature learning for multi-class Parkinson's disease classification. *Technol. Health Care* 26, 193–203. doi: 10.3233/THC-174548

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, H., Parikh, N. A., and He, L. (2018a). A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. *Front. Neurosci*. 12:491. doi: 10.3389/fnins.2018.00491

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* (Boston, MA: Springer), 305–312. doi: 10.1007/978-3-319-10443-0_39

Li, X., Dvornek, N. C., Papademetris, X., Zhuang, J., Staib, L. H., Ventola, P., et al. (2018b). "2-channel convolutional 3D deep neural network (2CC3D) for fMRI analysis: ASD classification and feature learning," in *2018 IEEE 15th*

International Symposium on Biomedical Imaging (ISBI 2018) (Washington, DC: IEEE), 1252–1255. doi: 10.1109/ISBI.2018.8363798

Li, Y., Meng, F., Shi, J., Initiative, A. D. N., et al. (2019). Learning using privileged information improves neuroimaging-based CAD of Alzheimer's disease: a comparative study. *Med. Biol. Eng. Comput*. 57, 1605–1616. doi: 10.1007/s11517-019-01974-3

Li, Z., Zhang, X., Müller, H., and Zhang, S. (2018c). Large-scale retrieval for medical image analytics: a comprehensive review. *Med. Image Anal*. 43, 66–84. doi: 10.1016/j.media.2017.09.007

Lian, C., Liu, M., Zhang, J., and Shen, D. (2018). Hierarchical fully convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Trans. Pattern Anal. Mach. Intell*. 42, 880–893. doi: 10.1109/TPAMI.2018.2889096

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

Liu, F., and Shen, C. (2014). Learning deep convolutional features for MRI based Alzheimer's disease classification. *arXiv* 1404.3366.

Liu, M., Cheng, D., Wang, K., Wang, Y., Initiative, A. D. N., et al. (2018a). Multimodality cascaded convolutional neural networks for Alzheimer's disease diagnosis. *Neuroinformatics* 16, 295–308. doi: 10.1007/s12021-018-9370-4

Liu, M., Cheng, D., and Yan, W. (2018b). 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. (2018c). 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, M., Zhang, J., Nie, D., Yap, P.-T., and Shen, D. (2018d). Anatomical landmark based deep feature representation for MR images in brain disease diagnosis. *IEEE J. Biomed. Health Inform*. 22, 1476–1485. doi: 10.1109/JBHI.2018.2791863

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, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M. F., Initiative, A. D. N., et al. (2018). Multiscale deep neural network based analysis of FDG-PET images for the early diagnosis of Alzheimer's disease. *Med. Image Anal*. 46, 26–34. doi: 10.1016/j.media.2018.02.002

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., et al. (2011). The parkinson progression marker initiative (PPMI). *Prog. Neurobiol*. 95, 629–635. doi: 10.1016/j.pneurobio.2011.09.005

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J.-M., Ramirez, J., and Castillo-Barnes, D. (2019). 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

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J. M., Ramirez, J., Castillo-Barnes, D., Salas-Gonzalez, D., et al. (2018). "Deep convolutional autoencoders vs PCA in a highly-unbalanced Parkinson's disease dataset: a DaTSCAN study," in *The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications* (San Sebastián: Springer), 47–56. doi: 10.1007/978-3-319-94120-2_5

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 *International Work-Conference on the Interplay Between Natural and Artificial Computation* (Corunna: Springer), 324–333. doi: 10.1007/978-3-319-59740-9_32

McDaniel, C., and Quinn, S. (2019). "Developing a graph convolution-based analysis pipeline for multi-modal neuroimage data: an application to Parkinson's Disease," in *Proceedings of the 18th Python in Science Conference (SciPy 2019)* (Austin, TX), 42–49. doi: 10.25080/Majora-7ddc1dd 1-006

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C., Jagust, W., et al. (2005). The Alzheimer's disease neuroimaging initiative. *Neuroimag. Clin*. 15, 869–877. doi: 10.1016/j.nic.2005.09.008

Oktay, O., Bai, W., Lee, M., Guerrero, R., Kamnitsas, K., Caballero, J., et al. (2016). "Multi-input cardiac image super-resolution using convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Athens: Springer), 246–254. doi: 10.1007/978-3-319-46726-9_29

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 *International Conference on Innovation in Medicine and Healthcare* (Puerto de la Cruz: Springer), 249–258. doi: 10.1007/978-3-319-39687-3_24

Pan, Y., Liu, M., Lian, C., Xia, Y., and Shen, D. (2019). "Disease-image specific generative adversarial network for brain disease diagnosis with incomplete multi-modal neuroimages," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen). doi: 10.1007/978-3-030-32248-9_16

Pan, Y., Liu, M., Lian, C., Zhou, T., Xia, Y., and Shen, D. (2018). "Synthesizing missing pet from mri with cycle-consistent generative adversarial networks for Alzheimer's disease diagnosis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Granada: Springer), 455–463. doi: 10.1007/978-3-030-00931-1_52

Pandya, M. D., Shah, P. D., and Jardosh, S. (2019). "Medical image diagnosis for disease detection: a deep learning approach," in *U-Healthcare Monitoring Systems* (Elsevier), 37–60. doi: 10.1016/B978-0-12-815370-3.00003-7

Parisot, S., Ktena, S. I., Ferrante, E., Lee, M., Moreno, R. G., Glocker, B., et al. (2017). "Spectral graph convolutions for population-based disease prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Quebec City, QC: Springer), 177–185. doi: 10.1007/978-3-319-66179-7_21

Patel, P., Aggarwal, P., and Gupta, A. (2016). "Classification of schizophrenia versus normal subjects using deep learning," in *Tenth Indian Conference on Computer Vision, Graphics and Image Processing* (Guwahati), 28. doi: 10.1145/3009977.3010050

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

Pinaya, W. H., Gadelha, A., Doyle, O. M., Noto, C., Zugman, A., Cordeiro, Q., et al. (2016). Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Sci. Rep*. 6, 1–9. doi: 10.1038/srep38897

Pinaya, W. H., Mechelli, A., and Sato, J. R. (2019). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: a large-scale multi-sample study. *Hum. Brain Mapp*. 40, 944–954. doi: 10.1002/hbm.24423

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

Poultney, C., Chopra, S., Cun, Y. L., and Ranzato, M. (2007). "Efficient learning of sparse representations with an energy-based model," in *Advances in Neural Information Processing Systems*, eds M. I. Jordan, Y. LeCun, and S. A. Solla (Vancouver, BC: ACM), 1137–1144. doi: 10.5555/2976456.2976599

Qi, J., and Tejedor, J. (2016). "Deep multi-view representation learning for multi-modal features of the schizophrenia and schizo-affective disorder," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (Shanghai), 952–956. doi: 10.1109/ICASSP.2016.7471816

Qureshi, M. N. I., Oh, J., and Lee, B. (2019). 3D-CNN based discrimination of schizophrenia using resting-state fMRI. *Artif. Intell. Med*. 98, 10–17. doi: 10.1016/j.artmed.2019.06.003

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* 1511.06434.

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Munich: Springer), 234–241. doi: 10.1007/978-3-319-24574-4_28

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* (Boston, MA: Springer), 520–527. doi: 10.1007/978-3-319-10404-1_65

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature* 323:533. doi: 10.1038/323533a0

Saeed, F., Eslami, T., Mirjalili, V., Fong, A., and Laird, A. (2019). ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data. *Front. Neuroinform*. 13:70. doi: 10.3389/fninf.2019.00070

Salakhutdinov, R. (2015). Learning deep generative models. *Annu. Rev. Stat. Appl.* 2, 361–385. doi: 10.1146/annurev-statistics-010814-020120

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.

Sarikaya, R., Hinton, G. E., and Deoras, A. (2014). Application of deep belief networks for natural language understanding. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 778–784. doi: 10.1109/TASLP.2014.2303296

Sarraf, S., Tofighi, G., et al. (2016). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *bioRxiv* 070441. doi: 10.1101/070441

Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U., and Langs, G. (2015). "Predicting semantic descriptions from medical images with convolutional neural networks," in *International Conference on Information Processing in Medical Imaging* (Springer), 437–448. doi: 10.1007/978-3-319-19992-4_34

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw*. 61, 85–117. doi: 10.1016/j.neunet.2014.09.003

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, L., Shi, J., Dong, Y., Ying, S., Peng, Y., Chen, L., et al. (2019a). An improved deep polynomial network algorithm for transcranial sonography-based diagnosis of Parkinson's disease. *Cogn. Comput*. 12, 553–562. doi: 10.1007/s12559-019-09691-7

Shen, T., Jiang, J., Lin, W., Ge, J., Wu, P., Zhou, Y., et al. (2019b). Use of overlapping group lasso sparse deep belief network to discriminate Parkinson's disease and normal control. *Front. Neurosci*. 13:396. doi: 10.3389/fnins.2019.00396

Shi, B., Chen, Y., Zhang, P., Smith, C. D., Liu, J., Initiative, A. D. N., et al. (2017a). Nonlinear feature transformation and deep fusion for Alzheimer's disease staging analysis. *Pattern Recogn*. 63, 487–498. doi: 10.1016/j.patcog.2016.09.032

Shi, J., Xue, Z., Dai, Y., Peng, B., Dong, Y., Zhang, Q., et al. (2018). Cascaded multi-column RVFL+ classifier for single-modal neuroimaging-based diagnosis of Parkinson's disease. *IEEE Trans. Biomed. Eng.* 66, 2362–2371. doi: 10.1109/TBME.2018.2889398

Shi, J., Zheng, X., Li, Y., Zhang, Q., and Ying, S. (2017b). 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

Shin, H.-C., Orton, M. R., Collins, D. J., Doran, S. J., and Leach, M. O. (2013). Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. *IEEE Trans. Pattern Analysis Mach. Intell*. 35, 1930–1943. doi: 10.1109/TPAMI.2012.277

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Singh, A., Sengupta, S., and Lakshminarayanan, V. (2020). Explainable deep learning models in medical image analysis. *arXiv* 2005.13799. doi: 10.3390/jimaging6060052

Singh, G., and Samavedham, L. (2015). Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: a case study on early-stage diagnosis of Parkinson disease. *J. Neurosci. Methods* 256, 30–40. doi: 10.1016/j.jneumeth.2015.08.011

Sivaranjini, S., and Sujatha, C. (2019). Deep learning based diagnosis of Parkinson's disease using convolutional neural network. *Multimed. Tools Appl*. 79, 15467–15479. doi: 10.1007/s11042-019-7469-8

Song, T.-A., Chowdhury, S. R., Yang, F., Jacobs, H., El Fakhri, G., Li, Q., et al. (2019). "Graph convolutional neural networks For Alzheimer's disease classification," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (Venice: IEEE), 414–417. doi: 10.1109/ISBI.2019.8759531

Srinivasagopalan, S., Barry, J., Gurupur, V., and Thankachan, S. (2019). A deep learning approach for diagnosing schizophrenic patients. *J. Exp. Theor. Artif. Intell*. 31, 1–14. doi: 10.1080/0952813X.2018.1563636

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. doi: 10.5555/2627435.2670313

Suk, H.-I., Lee, S.-W., Shen, D., and Alzheimers 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 Alzheimers Disease Neuroimaging Initiative. (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., Lee, S.-W., Shen, D., and Alzheimers Disease Neuroimaging Initiative. (2016). Deep sparse multi-task learning for feature selection in Alzheimer's disease diagnosis. *Brain Struct. Funct*. 221, 2569–2587. doi: 10.1007/s00429-015-1059-y

Sun, L., Shao, W., Wang, M., Zhang, D., and Liu, M. (2019a). High-order feature learning for multi-atlas based label fusion: application to brain segmentation with MRI. *IEEE Trans. Image Process*. 29, 2702–2713. doi: 10.1109/TIP.2019.2952079

Sun, L., Shao, W., Zhang, D., and Liu, M. (2019b). Anatomical attention guided deep networks for ROI segmentation of brain MR images. *IEEE Trans. Med. Imaging* 39, 2000–2012. doi: 10.1109/TMI.2019.2962792

Ulloa, A., Plis, S., and Calhoun, V. (2018). Improving classification rate of schizophrenia using a multimodal multi-layer perceptron model with structural and functional MR. *arXiv* 1804.04591.

Ulloa, A., Plis, S., Erhardt, E., and Calhoun, V. (2015). "Synthetic structural magnetic resonance image generator improves deep learning prediction of schizophrenia," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)* (Boston, MA), 1–6. doi: 10.1109/MLSP.2015.7324379

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

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* (Helsinki: ACM), 1096–1103. doi: 10.1145/1390156.1390294

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep learning for computer vision: a brief review. *Comput. Intell. Neurosci*. 2018:7068349. doi: 10.1155/2018/7068349

Wang, M., Hao, X., Huang, J., Shao, W., and Zhang, D. (2019a). Discovering network phenotype between genetic risk factors and disease status via diagnosis-aligned multi-modality regression method in Alzheimer's disease. *Bioinformatics* 35, 1948–1957. doi: 10.1093/bioinformatics/bty911

Wang, M., Lian, C., Yao, D., Zhang, D., Liu, M., and Shen, D. (2019b). Spatial-temporal dependency modeling and network hub detection for functional MRI analysis via convolutional-recurrent network. *IEEE Trans. Biomed. Eng*. 67, 2241–2252. doi: 10.1109/TBME.2019.2957921

Wang, M., Zhang, D., Huang, J., Shen, D., and Liu, M. (2018). "Low-rank representation for multi-center autism spectrum disorder identification," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2018* (Cham: Springer International Publishing), 647–654. doi: 10.1007/978-3-030-00928-1_73

Wang, M., Zhang, D., Huang, J., Yap, P.-T., Shen, D., and Liu, M. (2019c). Identifying autism spectrum disorder with multi-site fMRI via low-rank domain adaptation. *IEEE Trans. Med. Imaging* 39, 644–655. doi: 10.1109/TMI.2019.2933160

Wang, M., Zhang, D., Shen, D., and Liu, M. (2019d). Multi-task exclusive relationship learning for alzheimer's disease progression prediction with longitudinal data. *Med. Image Anal*. 53, 111–122. doi: 10.1016/j.media.2019.01.007

Wernick, M. N., Yang, Y., Brankov, J. G., Yourganov, G., and Strother, S. C. (2010). Machine learning in medical imaging. *IEEE Signal Process. Mag*. 27, 25–38. doi: 10.1109/MSP.2010.936730

Wu, G., Kim, M., Wang, Q., Gao, Y., Liao, S., and Shen, D. (2013). "Unsupervised deep feature learning for deformable registration of MR brain images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Shenzhen: Springer), 649–656. doi: 10.1007/978-3-642-40763-5_80

Wu, G., Shen, D., and Sabuncu, M. (2016). *Machine Learning and Medical Imaging*. Academic Press.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Yu, P. S. (2019). A comprehensive survey on graph neural networks. *arXiv* 1901.00596.

Yan, W., Calhoun, V., Song, M., Cui, Y., Yan, H., Liu, S., et al. (2019). Discriminating schizophrenia using recurrent neural network applied on time courses of multi-site fMRI data. *EBioMedicine* 47, 543–552. doi: 10.1016/j.ebiom.2019.08.023

Yao, D., Liu, M., Wang, M., Lian, C., Wei, J., Sun, L., et al. (2019). "Triplet graph convolutional network for multi-scale analysis of functional connectivity using functional MRI," in *International Workshop on Graph Learning in Medical Imaging* (Shenzhen: Springer), 70–78. doi: 10.1007/978-3-030-35817-4_9

Yu, S., Yue, G., Elazab, A., Song, X., Wang, T., and Lei, B. (2019). "Multi-scale graph convolutional network for mild cognitive impairment detection," in *International Workshop on Graph Learning in Medical Imaging* (Shenzhen: Springer), 79–87. doi: 10.1007/978-3-030-35817-4_10

Zeng, L. L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., et al. (2018). Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity MRI. *Ebiomedicine* 30, 74–85. doi: 10.1016/j.ebiom.2018.03.017

Zhang, D., Huang, J., Jie, B., Du, J., Tu, L., and Liu, M. (2018a). Ordinal pattern: a new descriptor for brain connectivity networks. *IEEE Trans. Med. Imaging* 37, 1711–1722. doi: 10.1109/TMI.2018.2798500

Zhang, X., He, L., Chen, K., Luo, Y., Zhou, J., and Wang, F. (2018b). "Multi-view graph convolutional network and its applications on neuroimage analysis for Parkinson's disease," in *AMIA Annual Symposium Proceedings*, Vol. 2018 (Washington, DC: American Medical Informatics Association), 1147.

Zhao, X., Zhou, F., Ou-Yang, L., Wang, T., and Lei, B. (2019). "Graph convolutional network analysis for mild cognitive impairment prediction," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)* (Venice: IEEE), 1598–601. doi: 10.1109/ISBI.2019.8759256

Zoph, B., and Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv* 1611.01578.

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 Neuroscience

ORIGINAL RESEARCH
published: 02 December 2020
doi: 10.3389/fnins.2020.591683

Fetal Cortical Plate Segmentation Using Fully Convolutional Networks With Multiple Plane Aggregation

Jinwoo Hong[1,2], Hyuk Jin Yun[2,3], Gilsoon Park[4], Seonggyu Kim[1], Cynthia T. Laurentys[2], Leticia C. Siqueira[2], Tomo Tarui[5,6], Caitlin K. Rollins[7], Cynthia M. Ortinau[8], P. Ellen Grant[2,3], Jong-Min Lee[4]* and Kiho Im[2,3]

[1] Department of Electronic Engineering, Hanyang University, Seoul, South Korea, [2] Fetal-Neonatal Neuroimaging and Developmental Science Center, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States, [3] Division of Newborn Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States, [4] Department of Biomedical Engineering, Hanyang University, Seoul, South Korea, [5] Mother Infant Research Institute, Tufts Medical Center, Tufts University School of Medicine, Boston, MA, United States, [6] Department of Pediatrics, Tufts Medical Center, Tufts University School of Medicine, Boston, MA, United States, [7] Department of Neurology, Boston Children's Hospital, Harvard Medical School, Boston, MA, United States, [8] Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, United States

OPEN ACCESS

Edited by:
Hyunjin Park,
Sungkyunkwan University,
South Korea

Reviewed by:
Hosung Kim,
University of Southern California, Los Angeles, United States
Laura Gui,
Université de Genève, Switzerland

***Correspondence:**
Jong-Min Lee
ljm@hanyang.ac.kr

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: 05 August 2020
Accepted: 04 November 2020
Published: 02 December 2020

Citation:
Hong J, Yun HJ, Park G, Kim S, Laurentys CT, Siqueira LC, Tarui T, Rollins CK, Ortinau CM, Grant PE, Lee J-M and Im K (2020) Fetal Cortical Plate Segmentation Using Fully Convolutional Networks With Multiple Plane Aggregation. Front. Neurosci. 14:591683. doi: 10.3389/fnins.2020.591683

Fetal magnetic resonance imaging (MRI) has the potential to advance our understanding of human brain development by providing quantitative information of cortical plate (CP) development *in vivo*. However, for a reliable quantitative analysis of cortical volume and sulcal folding, accurate and automated segmentation of the CP is crucial. In this study, we propose a fully convolutional neural network for the automatic segmentation of the CP. We developed a novel hybrid loss function to improve the segmentation accuracy and adopted multi-view (axial, coronal, and sagittal) aggregation with a test-time augmentation method to reduce errors using three-dimensional (3D) information and multiple predictions. We evaluated our proposed method using the ten-fold cross-validation of 52 fetal brain MR images (22.9–31.4 weeks of gestation). The proposed method obtained Dice coefficients of 0.907 ± 0.027 and 0.906 ± 0.031 as well as a mean surface distance error of 0.182 ± 0.058 mm and 0.185 ± 0.069 mm for the left and right, respectively. In addition, the left and right CP volumes, surface area, and global mean curvature generated by automatic segmentation showed a high correlation with the values generated by manual segmentation ($R^2 > 0.941$). We also demonstrated that the proposed hybrid loss function and the combination of multi-view aggregation and test-time augmentation significantly improved the CP segmentation accuracy. Our proposed segmentation method will be useful for the automatic and reliable quantification of the cortical structure in the fetal brain.

Keywords: deep learning, fetal brain, cortical plate, segmentation, hybrid loss, MRI

INTRODUCTION

A fundamental method for understanding brain development and disease is the quantitative analysis of magnetic resonance imaging (MRI) data, which requires preprocessing steps such as brain extraction, tissue segmentation (gray matter, white matter, and cerebrospinal fluid), and specific region-of-interest segmentation. Advances in MRI technology have enabled *in vivo* human fetal MRI studies to examine early brain development during the prenatal period. Among several quantitative indices of the human fetal brain, cortical volume and cortical folding patterns are crucial to the characterization and detection of abnormal brain development (Scott et al., 2011; Clouchoux et al., 2012; Im et al., 2013, 2017; Tarui et al., 2018; Ortinau et al., 2019; Yun et al., 2020a). For a reliable and sensitive analysis of its volume and surface folding patterns, accurate segmentation of the cortical plate (CP) is necessary. However, manual or semi-automatic segmentation has been used in previous studies which is a highly time-consuming and challenging task with high inter- and intra-rater variability. In addition, because fetal brains exhibit dramatic changes in size, cortical shape, cellular compartments, and image contrast at tissue boundaries, which vary with gestational age (GA) compared to child or adult brains, previous methods that were developed for the cortical gray matter segmentation of mature brains are not applicable to fetal brain segmentation.

Over the past decade, several algorithms for automatic CP segmentation from fetal MRI have been proposed. The expectation-maximization (EM) algorithm and atlas-based segmentation method have been employed for fetal brain tissue segmentation (Bach Cuadra et al., 2009; Habas et al., 2010; Serag et al., 2012; Wright et al., 2014). However, previous studies have reported results from a narrow GA range in a small number of subjects, and/or exhibited large errors [4–16 subjects, accuracy of CP segmentation measured by Dice coefficient = 0.63–0.84 and mean surface distance (MSD) error = 0.70–0.86 mm] (Bach Cuadra et al., 2009; Habas et al., 2010; Serag et al., 2012; Wright et al., 2014). The EM algorithm requires the precise estimation of a mixture of tissue probability using linear and non-linear registration between target images and a brain atlas. Likewise, atlas-based segmentation requires precise registration, including a non-linear approach between the target image and brain atlas. Fetal CPs have a very thin band-shaped structure, and the boundary of CPs is ambiguous owing to a low effective MRI resolution and the partial volume effect, which limits the accuracy of registration. Therefore, it may be difficult to accurately extract thin CPs from fetal MRI using the EM algorithm and atlas-based segmentation.

Recently, deep learning in the field of image segmentation has shown superior performance compared to traditional methods such as EM algorithm. Among various deep learning algorithms, the convolutional neural network (CNN) has been widely used for brain tissue and region segmentation in postnatal MRI data (Zhang et al., 2015; Kleesiek et al., 2016; Milletari et al., 2016; Ghafoorian et al., 2017; Chen et al., 2018; Kushibar et al., 2018; Wachinger et al., 2018; Alom et al., 2019; Guha Roy et al., 2019). Fetal CP segmentation methods based on MRI and ultrasound have been proposed using CNN (Khalili et al., 2019; Dou et al., 2020; Wyburd et al., 2020). One peer-reviewed MRI study proposed fetal brain tissue segmentation using a two-dimensional (2D) semantic CNN model that can segment seven brain tissues, including the CP (Khalili et al., 2019). However, the authors trained a CNN using the basic Dice loss, which maximize the Dice coefficient of segmentation. The basic Dice loss may not be optimal for relatively small areas in the multi-label segmentation problem, which may be a reason for the low accuracy of CP segmentation (Sudre et al., 2017; Wong et al., 2018). They obtained a CP segmentation accuracy that was relatively lower than the overall average Dice coefficient (CP: Dice coefficient = 0.835; Overall: Dice coefficient = 0.892) with a small number of fetal brain MRIs for a wide range of GA (12 fetuses from 22.9 to 34.6 weeks). Moreover, 3D information of the brain structures was not fully utilized in their methods, since they trained the network model using only coronal slices. To overcome the limitations in the previous methods, we propose an enhanced method for the automatic segmentation of the fetal CP using deep learning based on a large dataset of fetal brain MRIs. Our proposed method is focused on CP segmentation as our aim is to achieve the optimal accuracy of cortical volumes and surfaces. Numerous segmentation labels may require the complicated deep learning network and achieve inaccurate performance of CP segmentation. We propose a novel hybrid loss function and utilize a multi-view aggregation with test-time augmentation (MVT) approach to enhance the performance of CP segmentation. We adopt a focal Dice loss function, which is an exponential logarithmic Dice loss, to assign a large gradient to the less accurate labels (Wong et al., 2018). Our hybrid loss additionally includes a novel boundary Dice loss to accurately segment the CP boundary areas. In addition, the multi-view aggregation technique is used to enhance the segmentation accuracy by applying a 3D information to a 2D deep learning network. It combines three results from separate learning networks of 2D slices from three orthogonal planes (axial, coronal, and sagittal) to generate the final segmentation (Guha Roy et al., 2019; Jog et al., 2019; Estrada et al., 2020). The test-time augmentation (TTA) technique can obtain more robust prediction results using multiple predictions for a single input by applying the augmentation to test data, which is often used for the training phase in deep learning networks (Matsunaga et al., 2017; Jin et al., 2018). In this study, we applied both multi-view aggregation and TTA methods to obtain multiple results in each plane and to combine all results generated from the three planes. The hybrid loss was compared with the basic Dice loss, and MVT was compared with the results of the multi-view aggregation, TTA, and single view prediction. We hypothesized that MVT performs better than multi-view or TTA because it combines more segmentation results without changing the network and multi-view training structure. Furthermore, volume- and surface-based indices were extracted from both ground truth and automatic segmentation results and then compared to examine the reliability of brain measurements calculated from our segmentation.

MATERIALS AND METHODS

Dataset

The use of fetal MRIs was approved by the Institutional Review Boards at the Boston Children's Hospital (BCH) and Tufts Medical Center (TMC). Typically developing (TD) fetal MRIs were collected from subjects by recruitment, and retrospectively from clinical fetal MRIs performed to screen for abnormalities at BCH but found to be normal. Inclusion criteria for TD fetuses included no serious maternal medical conditions (nicotine or drug dependence, morbid obesity, cancer, diabetes, and gestational diabetes), maternal age between 18 and 45 years, fetal GA between 22 and 32 weeks GA. Exclusion criteria included multiple gestation pregnancies, dysmorphic features on ultrasound (US) examination, brain malformations, or brain lesions on US, other identified organ anomalies on US, known chromosomal abnormalities, known congenital infections and any abnormality on the fetal MRI. A total of 52 TD fetuses (22.9–31.4 weeks of pregnancy) were identified and used in this study. Fetal brain MRIs were acquired on a Siemens 3T Skyra scanner (BCH) or Phillips 1.5 T scanner (TMC) using a T2-weighted half-Fourier acquisition single-shot turbo spin-echo (HASTE) sequence with a 1-mm in-plane resolution, field of view (FOV) = 256 mm, TR = 1.5 s (BCH) or 12.5 s (TMC), TE = 120 ms (BCH) or 180 ms (TMC), and slice thickness = 2–4 mm. After localizing the fetal brain, the HASTE scans were acquired multiple times in different orthogonal orientations (a total of 3–10 scans) for reliable motion correction and the 3D reconstruction of fetal brain MRI.

Preprocessing

First, we performed preprocessing on fetal brain MRIs (Im et al., 2017; Tarui et al., 2018; Yun et al., 2019, 2020b). Using multiple scans of HASTE, a slice-to-volume registration technique was adopted to combine 2D slices of fetal brain MRIs to create a motion-corrected 3D volume (Kuklisova-Murgasova et al., 2012). We set the resolution of the reconstructed volume to a 0.75-mm isotropic voxel size. Because the size, position, and orientation of the reconstructed volumes vary for different fetuses, the reconstructed volumes were linearly registered to a fetal brain template using "FLIRT" in FSL and transformed to a standard coordinate space (Jenkinson et al., 2002; Serag et al., 2012). Then, the CP volume and whole inner volume of the CP were semi-automatically segmented into left and right based on the voxel intensities by two trained raters, and they were manually modified to obtain the final segmentation by a single person. The final segmentation from the semi-automatic approach was used as ground truth.

We performed additional processes on the registered MRI for better segmentation performance. First, we removed unnecessary non-brain voxels from the registered volume by multiplying them by the brain mask of the template. Second, the z-transformation was applied to normalize the intensity distribution across the entire MRI scan. Finally, the scanned image was cropped based on the size of the dilated template brain mask and the size of the 2D image of each axis plane, unified to a 128 × 128 2D slice by zero padding.

Network Architecture

The deep learning network architecture is shown in **Figure 1**. We configured the contracting (left side) path, expansive (right side) paths, and skip connections, similar to the U-Net (Ronneberger et al., 2015). The structure comprises repeated layers of the batch normalization (BN), exponential linear units (ELU), 3 × 3 zero-padded convolution, and a 2 × 2 max pooling with stride 2 (Ioffe and Szegedy, 2015; Clevert et al., 2016). Each network layer is divided into blocks based on the size of the feature map. Each block represents a structure in which the BN, ELU, and convolution layers are present in triplicate. The order of the layers in the block was composed of BN, ELU, and convolution by referring to the evaluation result of the previous study (He et al., 2016). Thirty-two feature maps were generated by convolution in the first block, and the number of feature maps doubled as the size of the block became smaller, finally generating 512 feature maps. In the expansive path, we extended the feature map of the lower feature map size block to the size of the higher size block using 3 × 3 transposed convolution. The extended feature map and the last feature map of a corresponding block on a contracting path of the same size were concatenated and used as inputs of repeated convolution. In the last layer, 1 × 1 convolution was employed to compress the desired number of labels from the 32 feature maps to 5 (including background), and softmax activation was applied to create a probability value for each label.

We additionally trained a 3D network to compare with the performance of the multi-view aggregation. The 3D network structure is basically the same with 2D network, and the 2D layers are simply changed to 3D layers (e.g., 2D convolution to 3D convolution). However, due to the limitation of the graphic processing unit (GPU) memory, the number of feature maps generated by convolution in the first block starts with eight, and the number of feature maps at the largest is 128.

Loss Function

Dice Loss

The Dice loss function was introduced in a previous medical image segmentation study (Milletari et al., 2016). The authors calculated the Dice loss using the Dice coefficient, which is an index used to evaluate the segmentation performance. For segmentation of the prostrate, the Dice loss exhibited superior performance to the re-weighted logistic loss. In this study, the Dice loss (L_{Dice}) was employed according to the following function:

$$L_{\mathrm{Dice}}(g,p) = 1 - \frac{1}{N_{\mathrm{l}}}\left(\sum\nolimits_{\mathrm{l}} \frac{2(\sum_{\mathrm{i}} g_{\mathrm{li}} p_{\mathrm{li}})) + \epsilon}{\sum_{\mathrm{i}}\left(g_{\mathrm{li}} + p_{\mathrm{li}}\right) + \epsilon}\right)$$

Here, i depicts the pixel location, l represents the label, and N_{l} is the total number of labels. p_{li} is the softmax probability calculated from the deep learning network, and g_{li} is the ground truth probability at location i and label l. ϵ is the smoothing term to prevent division by zero. The Dice coefficient of each label has a value between 0 and 1. The loss function (1– averaged Dice coefficient) is used for training.

Hybrid Loss

The Dice loss demonstrated its usefulness in the segmentation problem of medical images (Milletari et al., 2016; Guha Roy et al., 2019; Khalili et al., 2019). However, new losses that improve the Dice loss have recently been introduced (Sudre et al., 2017; Wong et al., 2018). The Dice loss is unfavorable for relatively

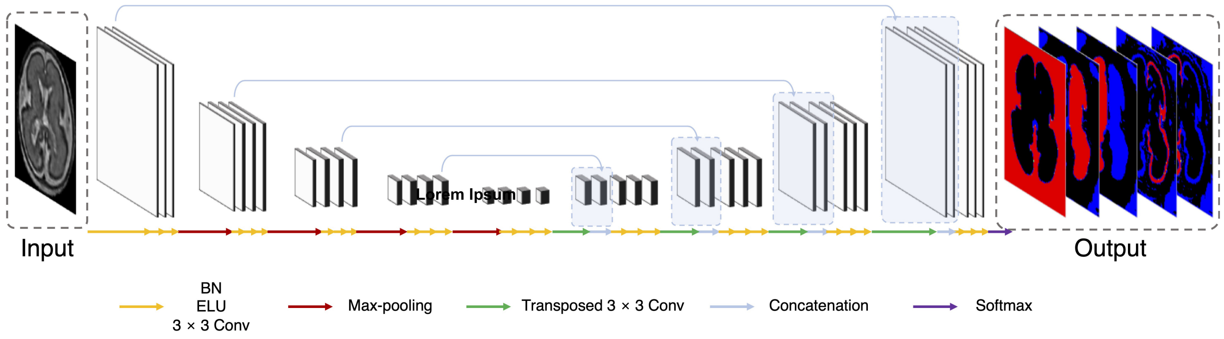


FIGURE 1 | Illustration of proposed network based on U-Net. Our network uses a 128 × 128 2D slice as the input and predicts the probability of five labels (background, left and right CP, and left and right inner volume of CP).

small structures, as misclassifying a few pixels can lead to a large reduction in the coefficient (Wong et al., 2018). Therefore, we adopted the logarithmic Dice loss (focal loss; L_{focal}), which focuses on less accurate labels (Wong et al., 2018):

$$L_{\text{focal}}(g, p) = \frac{1}{N_{\text{l}}}\left(\ln\left(\sum\nolimits_{\text{l}} \frac{2(\sum_{\text{i}} g_{\text{li}} p_{\text{li}})) + \epsilon}{\sum_{\text{i}}\left(g_{\text{li}} + p_{\text{li}}\right) + \epsilon}\right)^{\gamma}\right)$$

Here, γ dictates the non-linearities of the loss function. In this study, the optimum value of γ was 0.3 (Wong et al., 2018). This focal loss balances between structures that are easy and difficult to segment. Furthermore, we developed the boundary Dice loss to enhance the boundary segmentation accuracy. The Dice loss is effective at increasing the overall overlap between the ground truth and predictions; however, it lacks segmentation accuracy for boundary areas. Thus, to increase the weight of the boundary area, we calculated its Dice loss and added it to the loss for the entire area, which is called hybrid loss (L_{hyb}) in this paper.

$$L_{\text{hyb}}\left(g, p\right) = L_{\text{focal}}\left(g, p\right) + \lambda L_{\text{focal}}\left(g - g\ominus B, p - p\ominus B\right)$$

In the above equation, we use $\ominus$ to denote erosion; B is the erosion kernel (disk shape with diameter of 7), and λ is the weight for the boundary Dice loss. The boundary was detected through erosion and subtraction, and the Dice loss was calculated from the detected area and added to the whole-area Dice loss. The mixing weight λ was experimentally chosen by evaluating the Dice coefficient of the validation data for each λ in the range of 0.1–0.5; the best performance was obtained for $\lambda = 0.1$.

Aggregation

Multi-View Aggregation

Multi-view aggregation combines the predicted results in each orthogonal view, yielding a 3D regularization for errors occurring in 2D plane segmentation (Guha Roy et al., 2019). We trained a separate CNN for each of the three planes: axial, coronal, and sagittal. The predictions of each plane network were aggregated into the final segmentation map. The final segmentation map using multi-view aggregation (p_{mv}) was computed as follows:

$$p_{\text{mv}}\left(i\right) = \arg\max_{l}\left(p_{\text{axi}}\left(i, l\right) + p_{\text{cor}}\left(i, l\right) + p_{\text{sag}}\left(i, l\right)\right)$$

Here, $p_{\text{axi}}\left(i, l\right)$, $p_{\text{cor}}\left(i, l\right)$, and $p_{\text{sag}}\left(i, l\right)$ are the predicted four-dimensional probability arrays consisting of 3D of the voxel space and one dimension of the labels for axial, coronal, and sagittal planes, respectively. In the i-th voxel, the probabilities across the planes are summed and then a label with the highest probability is assigned as the final label. The predicted results for the axial and coronal planes (p_{axi} and p_{cor}) include 5 labels (background, left inner volume of CP, right inner volume of CP, left CP, and right CP), whereas result for sagittal plane (p_{sag}) contains only 3 labels (background, inner volume of CP, and CP) because there is no information on the left and right hemispheres in 2D sagittal view. Therefore, p_{sag} of the inner volume of CP is added to both probabilities of left and right inner volume of CP from other planes, and probability of CP is also added to both left and right. **Figure 2A** illustrates the multi-view aggregation.

Test-Time Augmentation

Test-time augmentation has been employed recently to improve the performance of various applications, including segmentation and classification (Matsunaga et al., 2017; Wang et al., 2019). The TTA technique was applied in the testing phase to improve the accuracy by creating various test results and combining these results. Ensemble of multiple prediction results for a single input can reduce prediction errors that may occur in a single prediction. We generated four outputs with artificially augmented inputs: original, horizontal flip, vertical flip, and horizontal/vertical flip (**Figure 2B**). In the case where slices are inverted left to right, the left and right sides of the output will be inverted from the original state. Therefore, when the left and right sides are inverted, an additional label inversion is applied to switch the left and right labels. For example, the final label map by the axial plane TTA (p_{TTA_axi}) is computed as follows:

$$p_{\text{sum\_axi}} = p_{\text{axi}} + T_{\text{h}}\left(p_{\text{axi}}\right) + T_{\text{v}}\left(p_{\text{axi}}\right) + T_{\text{hv}}\left(p_{\text{axi}}\right)$$

$$p_{\text{TTA\_axi}}\left(i\right) = \arg\max_{l}\left(p_{\text{sum}_{\text{axi}}}\left(i, l\right)\right)$$

Here, T_{h}, T_{v}, and T_{hv} are the horizontal flip, vertical flip, and horizontal/vertical flip transformation, respectively. This is similar to the multi-view aggregation in terms of combining

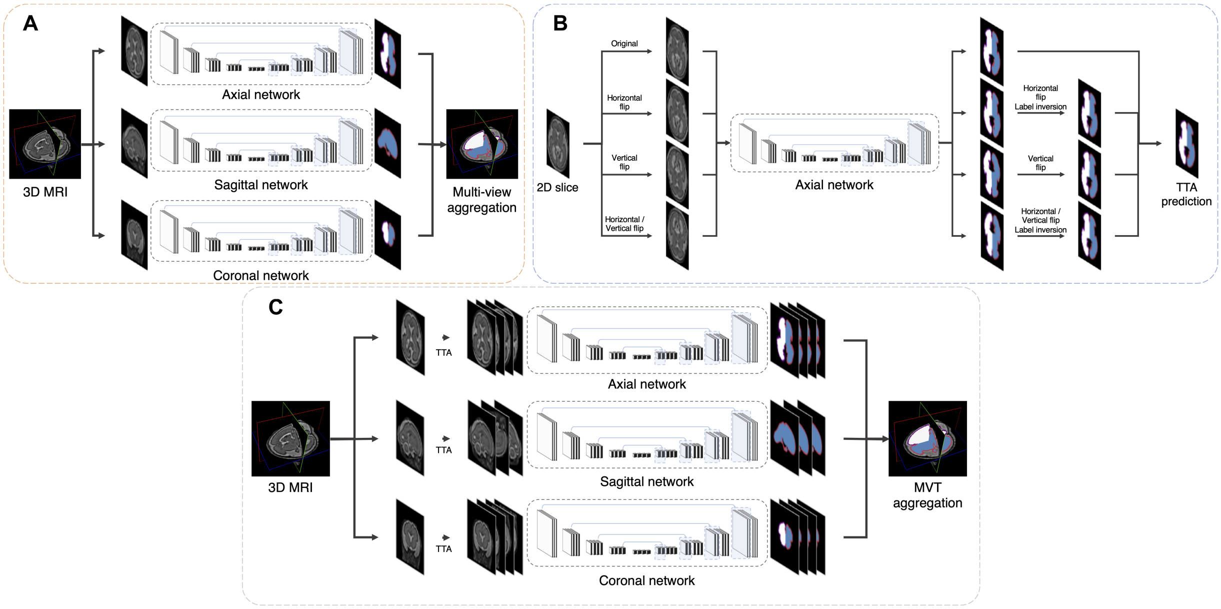


FIGURE 2 | Schematic representation of proposed segmentation procedure. **(A)** Multi-view aggregation combines segmentations from each trained model along three planes: coronal, sagittal, and axial. **(B)** TTA prediction synthesizes multiple segmentations by flip augmentation to generate a final segmentation map. **(C)** To enhance prediction accuracy, MVT aggregation is a combination of multi-view aggregation and TTA.

multiple results, whereas it differs from synthesizing multiple results in one view.

MVT Aggregation

MVT aggregation is a combination of 3D information from multi-view aggregation and ensemble of multiple predictions from TTA (**Figure 2C**). We applied TTA on each plane to obtain multiple results, and aggregated these results from each view to obtain the final result. In this study, the final label value on the i-th location $[p_{MVT}(i)]$ is computed as follows:

$$p_{\text{MVT}}(i) = \arg\max_{l} \left(p_{\text{sum\_axi}}(i, l) + p_{\text{sum\_cor}}(i, l) + p_{\text{sum\_sag}}(i, l)\right)$$

Here, $p_{\text{sum\_axi}}$, $p_{\text{sum\_cor}}$, and $p_{\text{sum\_sag}}$ are augmented prediction probability maps obtained from the axial, coronal, and sagittal planes, respectively. By increasing the number of prediction results used in the multi-view, more regularization effects are obtained than in the multi-view aggregation. A total of 11 (4 axial, 4 coronal, and 3 sagittal) prediction results were aggregated to generate the final 3D segmentation label map.

Training Strategy

Our model was tested with 52 fetuses using ten-fold cross-validation. Stratified sampling was used to match the GA distribution between training folds. 10% of the training samples selected through stratified sampling was used as a validation set. The hybrid loss described above was used for training, and deep learning was optimized using Adam (learning rate = 0.0001) (Kingma and Ba, 2015). For setting the optimal network weights in each fold, we monitored the Dice coefficient in the validation set in every epoch until there is no longer improvement of the Dice coefficient during the last 100 epochs using early stopping function. Then the network weights at the highest Dice coefficient in the validation set were stored as the optimal network. To increase the training dataset, data augmentation was applied. The augmentation parameters were vertical, horizontal, and vertical/horizontal flips. The type of data augmentation applied to the training phase was applied equally to the TTA prediction. For MVT aggregation, three networks of three orthogonal planes were trained. Although the three networks have the same structure, the number of the last outputs from the network of sagittal plane is different from those of axial and coronal planes, because the left and right hemispheres cannot be separated in sagittal plane.

Evaluation

The automatic segmentation performance was evaluated by the Dice coefficient used to measure the volume overlap and the MSD in order to quantify the boundary accuracy between the ground truth and the prediction segmentation map. The training of the network was based on 2D slices, whereas the proposed method evaluation was conducted in final 3D segmentation result. Furthermore, the CP volume and surface indices were measured and compared between the ground truth and automatically segmented volumes. To calculate the surface index, we adopted surface extraction procedure used in our previous studies (Im et al., 2017; Tarui et al., 2018; Yun et al., 2019, 2020b). Spatial smoothing was performed in the segmented inner volume of the CP using a 1.5 mm full width at half-maximum kernel to minimize noise. Using the smoothed inner volume of the CP, the hemispheric (left and right) triangular surface meshes of the inner CP boundary were automatically extracted by a function "isosurface" in MATLAB 2019b (MathWorks Inc., Natick, MA, United States). The surface models were geometrically smoothed using Freesurfer[1] to eliminate noise and small geometric changes. We calculated the CP volume based on the automatic segmentation result. Then, the surface area and global mean curvature (GMC) were calculated from the inner CP surface. The surface area was computed based on Voronoi region of each surface mesh vertex (Meyer et al., 2003). Mean curvature was defined as the angular deviation from each vertex (Meyer et al., 2003).

Statistical Analysis

We evaluated the effect of the loss and aggregation types on the automatic segmentation accuracy in four regions (left inner volume of CP, right inner volume of CP, left CP, and right CP) using the two-way repeated measure analysis of variance (ANOVA). Then, employing the *post hoc* test (Holm–Bonferroni method) for each effect, we determined which loss function and aggregation method performed best. The types of loss functions tested are basic Dice loss and hybrid loss, and the types of aggregation are MVT, multi-view, TTA$_{axi}$, TTA$_{cor}$, axi, and cor. The axi and cor denote the results obtained using only the original slice without any aggregation. There is no comparison for the sagittal plane since there is no information of the left and right hemispheres. TTA$_{axi}$ and TTA$_{cor}$ are obtained by applying TTA to the axial and coronal planes, respectively. Multi-view results are obtained from the combination of using only one result without TTA on the three planes, and MVT results from the combination of multiple results by applying TTA on all three planes. The numbers of segmentation aggregations are 11 (MVT), 3 (multi-view), 4 (TTA$_{axi}$), 4 (TTA$_{cor}$), 1 (axi), and 1 (cor). We used paired t-test to compare the performance between 2D multi-view network and 3D network. For direct comparison between the two networks, the same basic Dice loss was used without TTA. Subsequently, the similarities in the CP volume, surface area, and GMC between manual and automatic segmentation were evaluated using linear regression. Finally, we statistically evaluated whether the segmentation accuracies are associated with data properties, such as the subject age and imaging scanner. We evaluated GA-related changes of the Dice coefficient and MSD using the Pearson correlation analysis. Segmentation accuracies were statistically compared between different MR scanners (47 subjects from Siemens 3T at BCH vs. 5 subjects from Philips 1.5T at TMC) using a permutation test based on random resampling 10,000 times.

RESULTS

Effect of Loss Function

The repeated measure ANOVA test showed no difference between the Dice loss and hybrid loss in the inner volume of CP. However, the hybrid loss had a significantly higher segmentation accuracy (higher Dice coefficient and lower MSD) in the CP compared with the Dice loss (CP Dice coefficient [left, right]: $p = 0.027$, $p = 0.024$; CP MSD: $p = 0.024$, $p = 0.024$). The Dice coefficient and MSD for each loss are shown in **Table 1**. **Figure 3** shows an example of segmentation to verify the effect of hybrid loss.

Effect of Aggregation Method

The Dice coefficient and MSD for each aggregation method are shown in **Table 1**. In *post hoc* testing, axi and cor showed no statistical difference from each other in all four regions; however, they showed significantly increased accuracy when TTA was applied (TTA$_{axi}$ vs. axi and TTA$_{cor}$ vs. cor). There was no significant difference between TTA$_{axi}$ and TTA$_{cor}$. Multi-view aggregation exhibited a better performance than single plane-based TTA. Significantly large differences were found in most regions, except in the MSD of the right CP. Compared with other aggregation methods, the proposed MVT method yielded a significantly higher Dice coefficient in all *post hoc* tests. MVT

[1] https://surfer.nmr.mgh.harvard.edu

TABLE 1 | Statistical comparisons of segmentation performance obtained by different loss functions and aggregation methods.

		Loss		Aggregation					
		Hybrid	**Basic Dice**	**MVT**	**Multi-view**	**TTA$_{axi}$**	**TTA$_{cor}$**	**axi**	**cor**
Dice	in_L	0.978 ± 0.009	0.978 ± 0.009	0.980 ± 0.008	0.979 ± 0.008^{a}	0.978 ± 0.009^{a,b}	0.977 ± 0.009^{a,b}	0.977 ± 0.009^{a,b,c}	0.976 ± 0.009^{a,b,c,d}
	in_R	0.977 ± 0.011	0.977 ± 0.011	0.979 ± 0.011	0.978 ± 0.011^{a}	0.977 ± 0.012^{a,b}	0.977 ± 0.011^{a,b}	0.976 ± 0.011^{a,b,c,d}	0.976 ± 0.011^{a,b,c,d}
	CP_L	0.899 ± 0.027	0.885 ± 0.048*	0.907 ± 0.027	0.904 ± 0.027^{a}	0.897 ± 0.027^{a,b}	0.855 ± 0.126^{a,b}	0.894 ± 0.026^{a,b,c}	0.893 ± 0.029^{a,b,c}
	CP_R	0.898 ± 0.031	0.884 ± 0.050*	0.906 ± 0.031	0.902 ± 0.030^{a}	0.896 ± 0.032^{a,b}	0.896 ± 0.033^{a,b}	0.892 ± 0.031^{a,b,c,d}	0.851 ± 0.126^{a,b}
MSD	in_L	0.293 ± 0.092	0.293 ± 0.095	0.267 ± 0.092	0.277 ± 0.090^{a}	0.294 ± 0.097^{a,b}	0.299 ± 0.099^{a,b}	0.308 ± 0.097^{a,b,c}	0.312 ± 0.096^{a,b,c,d}
	in_R	0.300 ± 0.112	0.297 ± 0.110	0.271 ± 0.110	0.282 ± 0.107^{a}	0.299 ± 0.118^{a,b}	0.303 ± 0.116^{a,b}	0.318 ± 0.115^{a,b,c,d}	0.321 ± 0.108^{a,b,c,d}
	CP_L	0.199 ± 0.059	0.544 ± 1.064*	0.188 ± 0.060	0.190 ± 0.058	0.199 ± 0.060^{a,b}	1.229 ± 3.178	0.209 ± 0.060^{a,b,c}	0.213 ± 0.064^{a,b,c}
	CP_R	0.202 ± 0.070	0.551 ± 1.078*	0.186 ± 0.069	0.204 ± 0.077^{a}	0.203 ± 0.073^{a}	0.205 ± 0.073^{a}	0.215 ± 0.072^{a,c,d}	1.247 ± 3.192

Axi and cor denote one original slice result of axial and coronal planes, respectively. TTA$_{axi}$ and TTA$_{cor}$ aggregate four axial and four coronal view results, respectively. Multi-view aggregates three results (one axial, one coronal, and one sagittal). MVT combines 11 segmentation results (four axial, four coronal, and three sagittal) by applying TTA and multi-view simultaneously.

**Significantly different from hybrid loss; ^{a}significantly different from MVT; ^{b}significantly different from multi-view; ^{c}significantly different from TTA$_{axi}$; ^{d}significantly different from TTA$_{cor}$; all significant results: Holm–Bonferroni corrected $p < 0.05$.*

Data: mean ± standard deviation; in: inner volume of CP; L: left, R: right.

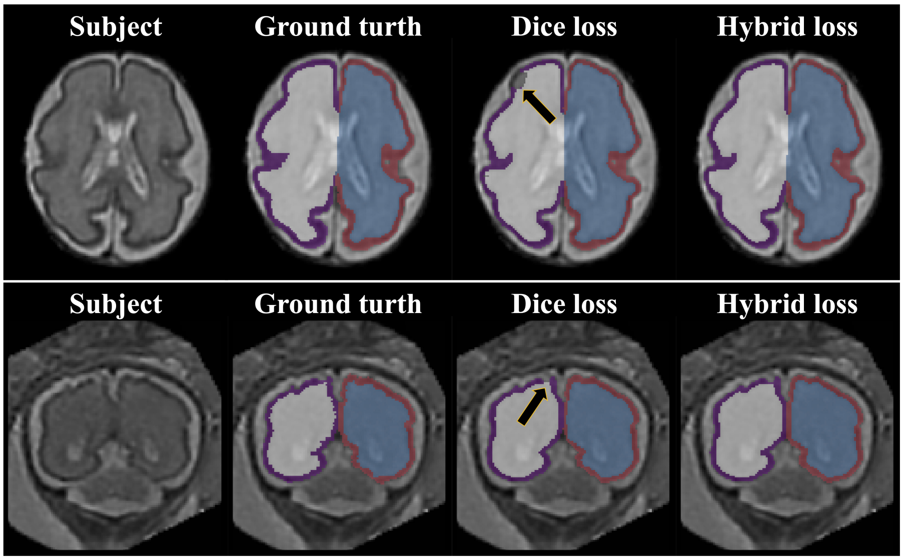


FIGURE 3 | Example of segmentation results with different loss function. The black arrows indicate the errors of segmentation when using the Dice loss. Since the loss for boundary was added, the proposed hybrid loss achieves more accurate segmentation results compared to the Dice loss.

also showed a significantly lower MSD than other methods in all comparisons except for those with multi-view and TTA_{cor} in the left CP and cor in the right CP. All statistical values of the comparisons among aggregation methods are shown in **Supplementary Tables 1–3**. **Figure 4** shows the example of segmentation to verify the effect of each aggregation method. For a visual comparison of the segmentation performance according to the aggregation method, box plots of both evaluation metrics are shown in **Figure 5**. Additionally, when compared to the 2D multi-view network, the 3D network obtained a significantly lower segmentation accuracy in both the Dice coefficient and MSD (see **Table 2**).

Volume and Surface Index Comparison

We evaluated similarity between the manual and our automatic segmentations in terms of the CP volume, area, and GMC of the inner CP surface. **Figure 6** shows the regression results between the indices obtained from the manual and automatic segmentations. The coefficient (β) of the linear regression is close to 1 in all indices, and it is statistically significant ($p < 0.0001$). An R^2 value of 0.94 or more is obtained for all indices. Therefore, the proposed method produced a very similar CP volume and surface indices when compared to manual segmentation.

Effects of Age and Scanner on Segmentation Performance

We evaluated the performance of the proposed method with respect to different GA and scanners. In terms of the MSD, for all regions, there were no significant changes of segmentation accuracy by GA (inner volume of CP [left, right]: $p = 0.113$, $p = 0.063$; CP: $p = 0.089$, $p = 0.055$). The Dice coefficient was significantly reduced with GA in the inner volume of CP (left: $p = 0.001$, right: $p = 0.002$). However, the correlations between the Dice coefficient and GA were not statistically significant in the left and right CP (left: $p = 0.055$, right: $p = 0.073$). **Figure 7** shows age-related trends of segmentation accuracy.

The accuracies obtained using automatic segmentation did not vary significantly across all regions between the two scanners (inner volume of CP Dice coefficient [left, right]: $p = 0.402$, $p = 0.406$; CP Dice coefficient: $p = 0.218$, $p = 0.239$; inner volume of CP MSD: $p = 0.603$, $p = 0.628$; CP MSD: $p = 0.384$, $p = 0.357$).

DISCUSSION

We developed a method to segment the CP of the fetal brain with high performance by employing the hybrid loss and MVT. The accuracy of the segmentation results obtained using our proposed method (Dice coefficient > 0.906, MSD < 0.185 mm) was superior to those using previous methods (Bach Cuadra et al., 2009; Habas et al., 2010; Serag et al., 2012; Wright et al., 2014; Khalili et al., 2019). Furthermore, the strong correlations of the volume-based index and surface-based indices between automatic and manual segmentation were found.

Hybrid Loss Function

We proposed a new hybrid loss to improve the segmentation accuracy at the boundary regions between tissues as well as the overall segmentation performance. Compared with the basic Dice loss, the hybrid loss showed significantly higher Dice coefficient and lower MSD (see **Table 1**). The proposed loss employed focal Dice loss in order to increase the overall performance, and focal boundary Dice loss in order to increase the boundary accuracy. In the multi-label segmentation problem, the adjustment of the segmentation weight between target labels in the network loss function is one of the primary factors affecting the performance (Sudre et al., 2017). The proposed method adopts a focal

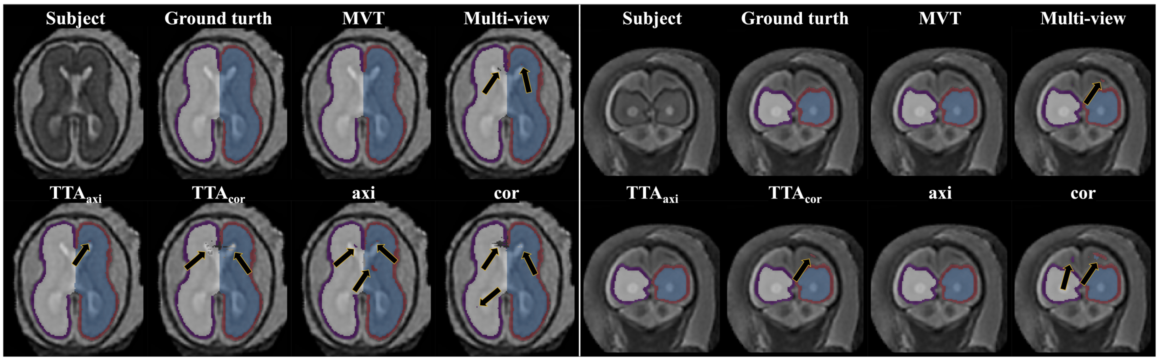


FIGURE 4 | Example of segmentation results with different aggregation methods. The black arrows indicate the errors of segmentation. The proposed MVT method effectively eliminated segmentation errors that remained even after using TTA or multi-view aggregation.

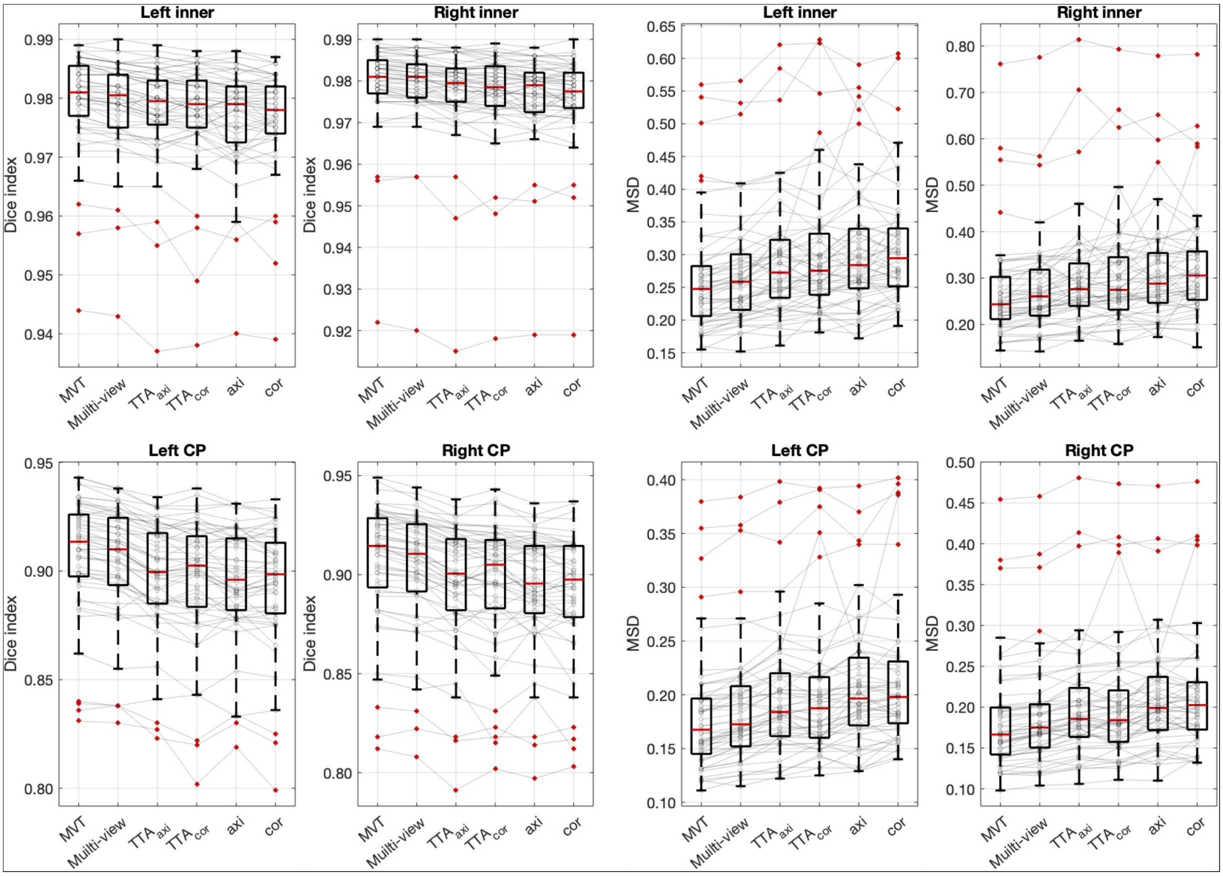


FIGURE 5 | Box plots of segmentation accuracy. The proposed method yields a significantly higher Dice coefficient and lower MSD compared with other methods. The gray line is the connection between the same subjects. *Post hoc* results are listed in **Table 1** and **Supplementary Tables 1–3**.

structure to adjust the segmentation weight without the need for a weight calculation process. The focal structure created using the logarithmic Dice loss assigns a larger gradient to lower-performance target labels (Wong et al., 2018). As we proposed, our result showed that the hybrid loss was more accurate than the basic Dice loss at the boundary area (**Figure 3**).

Recently, studies that employ boundary-related loss functions have been conducted (Schmidt and Boykov, 2012; Karimi and Salcudean, 2020). The Hausdorff distance (HD) loss was proposed to include the surface distance in the loss function (Karimi and Salcudean, 2020). However, the calculation process is complicated, and the weight compensation is difficult as the

TABLE 2 | Statistical comparisons of segmentation performance between 2D network with multi-view aggregation and 3D networks.

		2D multi-view	3D network	Paired *t*-test	
				t	*p*
Dice	in_L	0.979 ± 0.008	0.974 ± 0.010	8.352	0.0001
	in_R	0.978 ± 0.011	0.974 ± 0.011	8.563	0.0001
	CP_L	0.904 ± 0.028	0.819 ± 0.223	2.797	0.0073
	CP_R	0.901 ± 0.031	0.881 ± 0.033	12.822	0.0001
MSD	in_L	0.279 ± 0.092	0.369 ± 0.117	−8.067	0.0001
	in_R	0.283 ± 0.108	0.371 ± 0.137	−6.615	0.0001
	CP_L	0.190 ± 0.059	1.875 ± 5.565	−2.134	0.0377
	CP_R	0.217 ± 0.101	0.255 ± 0.081	−3.091	0.0032

Data, mean ± standard deviation; L, left; R, right.

range of values of the Dice and boundary loss vary. In this study, we proposed a morphological erosion-based boundary Dice loss which is simple and similar to the whole-area Dice loss, and the weight adjustment is straightforward as the range is the same as the whole-area Dice loss. An additional experiment was conducted to compare the segmentation performance between the HD loss (focal Dice loss + HD loss) and the hybrid loss proposed in this paper. There was no statistical difference between the two loss functions (paired *t*-test, CP Dice coefficient [left, right]: $p = 0.686$, $p = 0.544$; CP MSD: $p = 0.398$, $p = 0.243$). The proposed method has an advantage because it not only requires much simpler computation and weight control compared to the previous study (Karimi and Salcudean, 2020), but also shows a high segmentation performance.

MVT Aggregation

We propose the MVT aggregation, which combines multi-view aggregation and TTA. Compared to other aggregation methods, the proposed method showed significant increases in the Dice coefficient exhibited in all regions. The MSD significantly decreased in all regions except for the left CP of multi-view, the left CP of TTA$_{axi}$, and the right CP of cor. Our deep learning network did not fully utilize the 3D information of MRI as it was trained based on 2D slices. Therefore, to correct 2D results using 3D information, a multi-view aggregation was adopted, which synthesizes the results from networks of three orthogonal planes to generate a final 3D segmentation result. TTA was applied to improve the accuracy using various predicted segmentation maps. TTA improves the prediction accuracy by applying data augmentation to obtain multiple prediction results and ensemble them. As a result of the evaluation, we found that higher accuracies were obtained with a larger number of segmentation results (**Table 1** and **Supplementary Tables 1–3**). TTA results (TTA$_{axi}$ and TTA$_{cor}$) showed higher accuracies than those of one slice (axi and cor) (**Table 1** and **Supplementary Table 1**). The results demonstrate that multi-prediction by TTA can reduce errors that may occur in single prediction (axi and cor). Notably, multi-view results were more accurate than TTA (**Table 1** and **Supplementary Table 2**). For the final segmentation map, multi-view aggregation was corrected using three results from three planes, and it was more accurate than the TTA corrected with four results from one plane. This result indicates that 3D information from multi-view aggregation is more helpful for precise segmentation than the ensemble of the result using TTA. Also, multi-view aggregation outperformed the 3D network (**Table 2**). Although multi-view aggregation approach is based on a 2D network, it can reflect 3D information and utilize more training data, which may result in better segmentation performance compared to 3D network. MVT proposed in this study combines four results in the axial plane, four results in the coronal plane, and three results in the sagittal plane to finally produce the final segmentation with 11 prediction results. Therefore, the ensemble of multiple predictions using TTA was obtained, and at the same time, to generate more accurate segmentation results, the regularization of 3D information using multi-view aggregation was incorporated. The comparison of MVT with other approaches is shown in **Table 1** (**Supplementary Table 3**). **Figure 4** shows that as the level of the synthesis increases, the segmentation error decreases. It is shown that the error caused by prediction using only one slice can be corrected by TTA or multi-view, but more effectively by MVT.

Measurement of Volume- and Surface-Based Indices Using Automatic Segmentation

The accurate segmentation of brain regions is a fundamental step for the further analysis of brain morphometry using volume- and surface-based indices. The indices obtained from our segmentation method showed high correlations with the corresponding indices obtained from ground truth. When the CP volume is small, accurate results were obtained, whereas when the volume of the CP increased (>20 cc), the fitting accuracy decreased. This occurs because as the fetus grows and the brain size increases, the CP quickly becomes more complex and folded, increasing the difficulty of automatic segmentation. However, the actual average prediction errors remained low at values as small as 1.714 cc for the left CP and 2.308 cc for the right CP. Compared to the CP volume, regression models of surface indices showed higher correlations in the whole GA range between manual and automatic segmentation. Thus, our findings demonstrate that the proposed automatic segmentation method is reliable for further volume- and surface-based analyses.

Gestational Age and Scanner Effects on CP Segmentation

We used the Dice coefficient, MSD, and volume- and surface-based indices to evaluate the segmentation accuracy. Among them, in the fitting result for the CP volume, the accuracy of fitting tends to decrease as the volume of the CP increases. This trend is assumed to be related to the effect of the GA on the segmentation accuracy. The accuracy of CP segmentation exhibited a decreasing trend with an increase in GA, which is likely to result from the increasing complexity of the CP folding. However, the relationship between the GA and CP segmentation accuracy was not statistically significant. Upon measuring the accuracy for fetuses older than GA 30 weeks, the average Dice

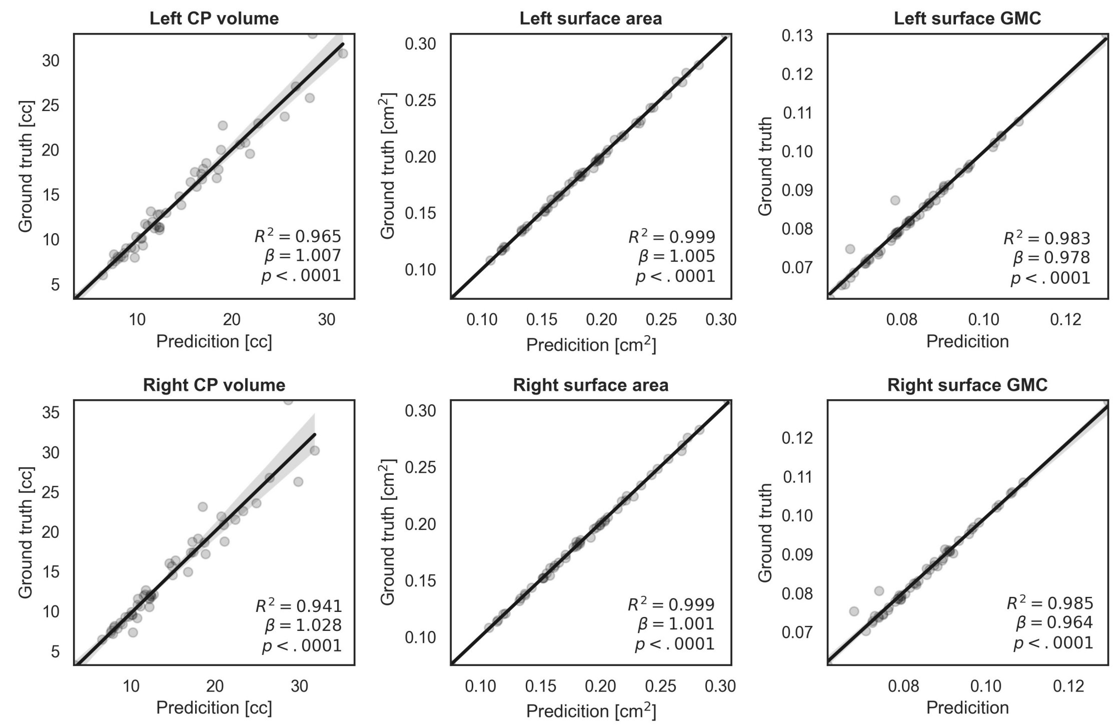


FIGURE 6 | Regression plots of volume, surface area, and surface GMC from ground truth and our automatic segmentation. The fitting result coefficient (β) was very close to unity in all indices in all regions.

coefficient and MSD were 0.967 and 0.323 mm, respectively, for the inner volume of CP, and 0.891 and 0.220 mm, respectively, for the CP. Hence, the proposed method demonstrated a high level of segmentation performance even in older fetuses. Additionally, **Supplementary Figure 1** shows local segmentation errors with different GA group. We divided the GA into three groups (22.9–25.3 [*n* = 19], 25.3–27.5 [17], 27.5–31.4 [16]), and showed examples of the segmentation errors for the subjects having the maximum, median, and minimum CP Dice coefficient in each GA group.

We performed permutation tests to verify whether there is any significant difference in the segmentation performance depending on the scanner. No statistical difference was found between scanners for all metrics, indicating that our results were not biased by the scanner effect.

Comparison With Other Methods

We propose a deep learning network for CP segmentation using MR images obtained from 52 fetuses. The proposed method obtained a Dice coefficient of 0.907 ± 0.027 and 0.906 ± 0.031, and an MSD of 0.182 ± 0.058 mm and 0.185 ± 0.069 mm for the left and right CP, respectively, using hybrid loss and MVT. Compared with other methods, we used a larger sample of the fetal dataset and varied the number of labels for segmentation. Therefore, it is difficult to compare the methods directly. Our proposed segmentation method was compared directly with a recent fetal CP segmentation deep learning model and indirectly with previous methods that used the EM algorithm and atlas-based segmentation. To the best of our knowledge, only two MRI studies and one ultrasound study have proposed the fetal CP segmentation method using deep learning (Khalili et al., 2019; Dou et al., 2020; Wyburd et al., 2020). Among them, our method was directly compared to one peer-reviewed study (Khalili et al., 2019). The authors applied a 2D U-Net with basic Dice loss to coronal MRI slices obtained from 12 fetuses, and a Dice coefficient of 0.835 and MSD of 0.307 mm were obtained for the CP volume (Khalili et al., 2019). When compared with our proposed deep learning model, the structure of the model was the same, but the loss function used for training was different and the MVT approach was not used.

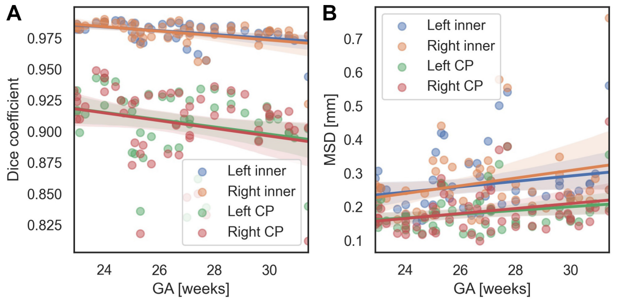


FIGURE 7 | Age-related trends of segmentation accuracy of the proposed method. **(A)** Dice coefficient. **(B)** MSD.

TABLE 3 | Cortical plate (CP) segmentation performance of the proposed method and other methods.

		Deep learning (direct)		**EM (indirect)**			**Atlas-based (indirect)**
		Proposed	**Khalili et al., 2019**	**Bach Cuadra et al., 2009**	**Habas et al., 2010**	**Wright et al., 2014**	**Serag et al., 2012**
No. subject (GA range)		52 (22.9–31.4)	52 (22.9–31.4)	4 (29–32)	14 (20.6–22.9)	16 (22.4–36.4)	15 (21.7–38.7)
Dice	CP_L	**0.907 ± 0.027**	0.894 ± 0.030	–	–	–	–
	CP_R	**0.906 ± 0.031**	0.811 ± 0.251	–	–	–	–
	CP	**0.907 ± 0.029**	0.852 ± 0.141	0.625 ± 0.038	0.82 ± 0.02	–	0.84 ± 0.06
MSD	CP_L	**0.182 ± 0.058**	0.212 ± 0.064	–	–	–	–
	CP_R	**0.185 ± 0.069**	2.277 ± 6.388	–	–	––	–
	CP	**0.184 ± 0.063**	1.245 ± 3.226	0.697 ± 0.079	–	0.864 ± 0.141	–

For direct comparison with the previous deep learning method of (Khalili et al., 2019), we implemented their method and applied it to our dataset. The Dice coefficient and MSD of EM and atlas-based methods were taken from their published papers for indirect comparison. The proposed method shows a higher Dice coefficient and a lower MSD when compared to previous studies either directly or indirectly.

Data, mean ± standard deviation; L, left; R, right. Bold values indicate the results of our proposed method that show the best performance.

Therefore, of the results in this paper, the result obtained using basic Dice loss in the network for coronal slices can be considered to result from the method of the prior study (CP Dice coefficient [left, right] = 0.894 ± 0.030, 0.811 ± 0.251; CP MSD = 0.212 ± 0.064 mm, 2.277 ± 6.388 mm). The proposed method showed a significantly higher segmentation accuracy using hybrid loss and MVT compared to the prior deep learning method (Khalili et al., 2019) (CP Dice coefficient [left, right]: $p < 0.0001$, $p = 0.009$; CP MSD : $p < 0.0001$, $p = 0.022$). The results obtained by the EM algorithm were as follows: (Bach Cuadra et al., 2009): 4 subjects; 29–32 weeks GA; Dice coefficient = 0.63 ± 0.04; MSD = 0.70 ± 0.08 mm, (Habas et al., 2010): 14 subjects; 20.57–22.86 weeks GA; Dice coefficient = 0.82 ± 0.02, (Wright et al., 2014): 16 subjects; 22.4–36.4 weeks GA; MSD = 0.86 ± 0.14 mm. The atlas-based segmentation method reported a Dice coefficient of 0.84 ± 0.06 for CP using MRI data from 15 fetuses (21.7–38.7 weeks GA) (Serag et al., 2012). Detailed results are shown in **Table 3**. Our method shows a better performance in terms of both the Dice coefficient and MSD when directly or indirectly compared to previous methods. The GA range of fetal subjects included in our study is narrower compared to some of the previous studies (Serag et al., 2012; Wright et al., 2014), which may result in higher accuracy as the older fetal brain MRI scans with complex folding are more difficult for CP segmentation. However, compared with the results obtained in our study, those studies utilized very few fetal MRI scans (≤ 16) and showed considerable differences in the Dice coefficient and MSD. Moreover, we found no significant correlations between the GA and CP segmentation accuracy, and obtained high accuracies even for fetuses over 30 weeks GA, as described above. Therefore, the narrow GA range in our study was not a bias causing the high accuracy. The previous deep learning study employed the basic Dice loss in multi-label segmentation, and showed relatively poor performance in small volume labels (Sudre et al., 2017; Wong et al., 2018). Although the authors applied several augmentation methods to increase the amount of training data in deep learning, they did not include the correction achieved by multiple predictions. The higher accuracy obtained in our method may be attributed to the inclusion of a loss function suitable for multi-label segmentation and correction by multiple predictions using MVT. The relatively low performance of the EM algorithm and atlas-based segmentation may be due to the registration quality as the brain template created by combining multiple images is blurred compared to individual images. It is not easy to obtain an accurate registration of the brain template to a target subject image, even with non-linear transformation. Furthermore, the partial volume effect of the CP boundary owing to the limited fetal MRI resolution and motion decreases the accuracy with which the likelihood probability of the EM algorithm and the registration accuracy of the atlas-based method can be estimated. The proposed method used only linear registration to unify the size of input images. Unlike previous methods, deep learning is free of registration effects because it does not need to accurately match any prior information. Furthermore, the inaccuracy that results from the partial volume effect may also be sufficiently trained by deep learning to enable a similar segmentation, as is possible with the ground truth. The proposed deep learning network exhibits a higher segmentation performance using hybrid loss and MVT than other methods.

Limitations

Despite the accurate CP segmentation with MVT and hybrid loss, there are some limitations to the proposed method. First, because the folding pattern of the fetal brain changes dynamically and becomes more complex as gestation progresses, a decreasing trend was observed in the CP segmentation accuracy with age although it was not statistically significant. Therefore, to improve the segmentation accuracy, it is necessary to include a larger number of fetuses above 30 weeks GA. Second, the proposed model did not include cerebrospinal fluid (CSF). In particular, the segmentation of deep sulcal CSF is essential for precise outer CP surface extraction, which enables the further analysis of cortical measures, such as cortical thickness. However, because of the limited resolution of fetal brain MRI scans and the partial volume effect of CSF in narrow deep sulcal regions, the manual segmentation of CSF in these regions is highly challenging. Although CSF segmentation was included in previous studies (Wright et al., 2014; Khalili et al., 2019), it has not been designed to extract deep sulcal CSF. In future studies, we will carefully delineate fetal CSF regions and train them to develop an automatic method for CSF segmentation.

CONCLUSION

The proposed method segments the fetal CP providing highly accurate measurements of CP volume and the highly accurate surface reconstruction of the CP. The hybrid loss and MVT show a significant increase in accuracy compared to the basic Dice loss and other aggregation methods. Although most of our comparisons were performed indirectly, the proposed method showed better fetal CP segmentation performance than other methods. Likewise, the comparisons of CP volume and surface indices between prediction and ground truth showed high similarity. Our results indicate that our proposed automatic segmentation method is useful for performing an accurate quantitative cortical structural analysis in the human fetal brain. The developed automatic segmentation is more reproducible than manual segmentation as it is not affected by inter- and intra-rater variability, and it has a short computation time.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are not readily available because the fetal MRIs used in this study was not available publically. Requests to access the datasets should be directed to KI, kiho.im@childrens.harvard.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Boards at the Boston Children's Hospital (BCH) and Tufts Medical Center (TMC). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

KI, HJY, and J-ML designed the main idea and directed the overall analysis. JH and HJY developed the algorithm and carried out the data processing and experiments. GP, SK, CL, LS, TT, CR, CO, and PG assisted with the data collection and result interpretation. JH, HJY, KI, and J-ML wrote the manuscript with input from all authors.

FUNDING

This research was 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 (grant number: HI19C0755); Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. 2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University)]; Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (NIH) (R21HD094130, U01HD087211, R01HD100009, and K23HD079605); National Institute of Neurological Disorders and Stroke of the National Institutes of Health (R01NS114087); American Heart Association (19IPLOI34660336); National Institutes of Health/National Heart, Lung, and Blood Institute (K23HL141602); National Institute of Neurological Disorders and Stroke (K23-NS101120); National Institute of Biomedical Imaging and Bioengineering (R01EB017337); and the Susan Saltonstall Foundation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2020.591683/full#supplementary-material

REFERENCES

Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., and Asari, V. K. (2019). Recurrent residual U-Net for medical image segmentation. *J. Med. Imaging* 6:014006. doi: 10.1117/1.jmi.6.1.014006

Bach Cuadra, M., Schaer, M., Andre, A., Guibaud, L., Eliez, S., and Thiran, J.-P. (2009). Brain tissue segmentation of fetal MR images. *Int. Conf. on Med. Image Comput. and Comput. Assist. Interv.* 2009, 1–9.

Chen, H., Dou, Q., Yu, L., Qin, J., and Heng, P. A. (2018). VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. *Neuroimage* 170, 446–455. doi: 10.1016/j.neuroimage.2017.04.041

Clevert, D. A., Unterthiner, T., and Hochreiter, S. (2016). "Fast and accurate deep network learning by exponential linear units (ELUs)," in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings. 2016*, San Juan.

Clouchoux, C., Kudelski, D., Gholipour, A., Warfield, S. K., Viseur, S., Bouyssi-Kobar, M., et al. (2012). Quantitative in vivo MRI measurement of cortical development in the fetus. *Brain Struct. Funct.* 217, 127–139. doi: 10.1007/s00429-011-0325-x

Dou, H., Karimi, D., Rollins, C. K., Ortinau, C. M., Vasung, L., Velasco-Annis, C., et al. (2020). *A Deep Attentive Convolutional Neural Network for Automatic Cortical Plate Segmentation in Fetal MRI*. Available at: https://github.com/wulalago/FetalCPSeg (accessed June 29, 2020).

Estrada, S., Lu, R., Conjeti, S., Orozco-Ruiz, X., Panos-Willuhn, J., Breteler, M. M. B., et al. (2020). FatSegNet: a fully automated deep learning pipeline for adipose tissue segmentation on abdominal dixon MRI. *Magn. Reson. Med.* 83, 1471–1483. doi: 10.1002/mrm.28022

Ghafoorian, M., Karssemeijer, N., Heskes, T., Van Uden, I. W. M., Sanchez, C. I., Litjens, G., et al. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Sci. Rep.* 7:5110. doi: 10.1038/s41598-017-05300-5

Guha Roy, A., Conjeti, S., Navab, N., and Wachinger, C. (2019). QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. *Neuroimage* 186, 713–727. doi: 10.1016/j.neuroimage.2018.11.042

Habas, P. A., Kim, K., Rousseau, F., Glenn, O. A., Barkovich, A. J., and Studholme, C. (2010). Atlas-based segmentation of developing tissues in the human brain with quantitative validation in young fetuses. *Hum. Brain Mapp.* 31, 1348–1358. doi: 10.1002/hbm.20935

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Identity mappings in deep residual networks," in *Lecture Notes in Computer Science (including subseries Lecture*

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), eds B. Leibe, J. Matas, N. Sebe, and M. Welling, (Berlin: Springer Verlag), 630–645. doi: 10.1007/978-3-319-46493-0_38

Im, K., Guimaraes, A., Kim, Y., Cottrill, E., Gagoski, B., Rollins, C., et al. (2017). Quantitative folding pattern analysis of early primary sulci in human fetuses with brain abnormalities. *Am. J. Neuroradiol.* 38, 1449–1455. doi: 10.3174/ajnr.A5217

Im, K., Pienaar, R., Paldino, M. J., Gaab, N., Galaburda, A. M., and Grant, P. E. (2013). Quantification and discrimination of abnormal sulcal patterns in polymicrogyria. *Cereb. Cortex* 23, 3007–3015. 10.1093/cercor/bhs292

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning, ICML 2015*, France: International Machine Learning Society (IMLS), 448–456.

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

Jin, H., Li, Z., Tong, R., and Lin, L. (2018). A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection. *Med. Phys.* 45, 2097–2107. doi: 10.1002/mp.12846

Jog, A., Grant, P. E., Jacobson, J. L., van der Kouwe, A., Meintjes, E. M., Fischl, B., et al. (2019). Fast infant MRI skullstripping with multiview 2D convolutional neural networks. *arXiv* [Preprint], Available at: http://arxiv.org/abs/1904.12101 (accessed February 27, 2020).

Karimi, D., and Salcudean, S. E. (2020). Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Trans. Med. Imaging* 39, 499–513. doi: 10.1109/TMI.2019.2930068

Khalili, N., Lessmann, N., Turk, E., Claessens, N., de Heus, R., Kolk, T., et al. (2019). Automatic brain tissue segmentation in fetal MRI using convolutional neural networks. *Magn. Reson. Imaging* 64, 77–89. doi: 10.1016/j.mri.2019.05.020

Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *3rd Int. Conf. on Learn. Represent., ICLR 2015 - Conf. Track Proc*, San Diego, CA.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., et al. (2016). Deep MRI brain extraction: a 3D convolutional neural network for skull stripping. *Neuroimage* 129, 460–469. doi: 10.1016/j.neuroimage.2016.01.024

Kuklisova-Murgasova, M., Quaghebeur, G., Rutherford, M. A., Hajnal, J. V., and Schnabel, J. A. (2012). Reconstruction of fetal brain MRI with intensity matching and complete outlier removal. *Med. Image Anal.* 16, 1550–1564. doi: 10.1016/j.media.2012.07.004

Kushibar, K., Valverde, S., González-Villà, S., Bernal, J., Cabezas, M., Oliver, A., et al. (2018). Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. *Med. Image Anal.* 48, 177–186. doi: 10.1016/j.media.2018.06.006

Matsunaga, K., Hamada, A., Minagawa, A., and Koga, H. (2017). Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv* [Preprint], Available at: http://arxiv.org/abs/1703.03108 (accessed March 5, 2020).

Meyer, M., Desbrun, M., Schröder, P., and Barr, A. H. (2003). "Discrete differential-geometry operators for triangulated 2-manifolds bt - visualization and mathematics III," in *Visualization and Mathematics III. Mathematics and Visualization*, eds H. C. Hege, and K. Polthier, (Berlin: Springer), doi: 10.1007/978-3-662-05105-4_2

Milletari, F., Navab, N., and Ahmadi, S. A. (2016). "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Proc. - 2016 4th Int. Conf. on 3D Vision, 3DV 2016*, Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc, 565–571. doi: 10.1109/3DV.2016.79

Ortinau, C. M., Rollins, C. K., Gholipour, A., Yun, H. J., Marshall, M., Gagoski, B., et al. (2019). Early-emerging sulcal patterns are atypical in fetuses with congenital heart disease. *Cereb. Cortex* 29, 3605–3616. doi: 10.1093/cercor/bhy235

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds B. Leibe, J. Matas, N. Sebe, and M. Welling, (Berlin: Springer Verlag), 234–241. doi: 10.1007/978-3-319-24574-4_28

Schmidt, F. R., and Boykov, Y. (2012). "Hausdorff distance constraint for multi-surface segmentation," in *Lecture Notes in Computer Science. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds B. Leibe, J. Matas, N. Sebe, and M. Welling, (Berlin: Springer Verlag), 598–611. doi: 10.1007/978-3-642-33718-5_43

Scott, J. A., Habas, P. A., Kim, K., Rajagopalan, V., Hamzelou, K. S., Corbett-Detig, J. M., et al. (2011). Growth trajectories of the human fetal brain tissues estimated from 3D reconstructed in utero MRI. *Int. J. Dev. Neurosci.* 29, 529–536. doi: 10.1016/j.ijdevneu.2011.04.001

Serag, A., Edwards, A. D., Hajnal, J. V., Counsell, S. J., Boardman, J. P., and Rueckert, D. (2012). A multi-channel 4D probabilistic atlas of the developing brain: application to fetuses and neonates. *Spec. Issue Ann. Br. Mach. Vis. Assoc.* 2012, 1–14.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Jorge Cardoso, M. (2017). "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Lecture Notes in Computer Science. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds B. Leibe, J. Matas, N. Sebe, and M. Welling, (Berlin: Springer Verlag), 240–248. doi: 10.1007/978-3-319-67558-9_28

Tarui, T., Madan, N., Farhat, N., Kitano, R., Tanritanir, A. C., Graham, G., et al. (2018). Disorganized patterns of sulcal position in fetal brains with agenesis of corpus callosum. *Cereb. Cortex* 28, 3192–3203. doi: 10.1093/cercor/bhx191

Wachinger, C., Reuter, M., and Klein, T. (2018). DeepNAT: deep convolutional neural network for segmenting neuroanatomy. *Neuroimage* 170, 434–445. doi: 10.1016/j.neuroimage.2017.02.035

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. (2019). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing* 338, 34–45. doi: 10.1016/j.neucom.2019.01.103

Wong, K. C. L., Moradi, M., Tang, H., and Syeda-Mahmood, T. (2018). "3D segmentation with exponential logarithmic loss for highly unbalanced object sizes," in *Lecture Notes in Computer Science. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds B. Leibe, J. Matas, N. Sebe, and M. Welling, (Berlin: Springer Verlag), 612–619. doi: 10.1007/978-3-030-00931-1_70

Wright, R., Kyriakopoulou, V., Ledig, C., Rutherford, M. A., Hajnal, J. V., Rueckert, D., et al. (2014). Automatic quantification of normal cortical folding patterns from fetal brain MRI. *Neuroimage* 91, 21–32. doi: 10.1016/j.neuroimage.2014.01.034

Wyburd, M. K., Jenkinson, M., and Namburete, A. I. L. (2020). "Cortical plate segmentation using CNNs in 3D fetal ultrasound," in *Communications in Computer and Information Science*, eds Papież, W. Bartłomiej, Namburete, I. L. Ana, Yaqub, Mohammad, et al. (Berlin: Springer), 56–68. doi: 10.1007/978-3-030-52791-4_5

Yun, H. J., Chung, A. W., Vasung, L., Yang, E., Tarui, T., Rollins, C. K., et al. (2019). Automatic labeling of cortical sulci for the human fetal brain based on spatio-temporal information of gyrification. *Neuroimage* 188, 473–482. doi: 10.1016/j.neuroimage.2018.12.023

Yun, H. J., Perez, J. D. R., Sosa, P., Valdés, J. A., Madan, N., Kitano, R., et al. (2020a). Regional alterations in cortical sulcal depth in living fetuses with down syndrome. *Cereb. Cortex* bhaa255. doi: 10.1093/cercor/bhaa255

Yun, H. J., Vasung, L., Tarui, T., Rollins, C. K., Ortinau, C. M., Grant, P. E., et al. (2020b). Temporal patterns of emergence and spatial distribution of sulcal pits during fetal life. *Cereb. Cortex* 30, 4257–4268. doi: 10.1093/cercor/bhaa053

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., et al. (2015). Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. *Neuroimage* 108, 214–224. doi: 10.1016/j.neuroimage.2014.12.061

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 Neuroscience

ORIGINAL RESEARCH
published: 07 July 2021
doi: 10.3389/fnins.2021.652987

Three-Dimensional Convolutional Autoencoder Extracts Features of Structural Brain Images With a "Diagnostic Label-Free" Approach: Application to Schizophrenia Datasets

*Hiroyuki Yamaguchi[1,2], Yuki Hashimoto[1], Genichi Sugihara[3], Jun Miyata[4], Toshiya Murai[4], Hidehiko Takahashi[3], Manabu Honda[1], Akitoyo Hishimoto[2] and Yuichi Yamashita[1]**

[1] Department of Information Medicine, National Center of Neurology and Psychiatry, National Institute of Neuroscience, Tokyo, Japan, [2] Department of Psychiatry, School of Medicine, Yokohama City University, Yokohama, Japan, [3] Department of Psychiatry and Behavioral Sciences, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan, [4] Department of Psychiatry, Graduate School of Medicine, Kyoto University, Kyoto, Japan

There has been increasing interest in performing psychiatric brain imaging studies using deep learning. However, most studies in this field disregard three-dimensional (3D) spatial information and targeted disease discrimination, without considering the genetic and clinical heterogeneity of psychiatric disorders. The purpose of this study was to investigate the efficacy of a 3D convolutional autoencoder (3D-CAE) for extracting features related to psychiatric disorders without diagnostic labels. The network was trained using a Kyoto University dataset including 82 patients with schizophrenia (SZ) and 90 healthy subjects (HS) and was evaluated using Center for Biomedical Research Excellence (COBRE) datasets, including 71 SZ patients and 71 HS. We created 16 3D-CAE models with different channels and convolutions to explore the effective range of hyperparameters for psychiatric brain imaging. The number of blocks containing two convolutional layers and one pooling layer was set, ranging from 1 block to 4 blocks. The number of channels in the extraction layer varied from 1, 4, 16, and 32 channels. The proposed 3D-CAEs were successfully reproduced into 3D structural magnetic resonance imaging (MRI) scans with sufficiently low errors. In addition, the features extracted using 3D-CAE retained the relation to clinical information. We explored the appropriate hyperparameter range of 3D-CAE, and it was suggested that a model with 3 blocks may be related to extracting features for predicting the dose of medication and symptom severity in schizophrenia.

Keywords: deep learning, machine learning, neuroimaging, schizophrenia, structural MRI, convolutional autoencoder, diagnostic label

OPEN ACCESS

Edited by:
Albert Yang,
National Yang-Ming University, Taiwan

Reviewed by:
Adil Bashir,
Auburn University, United States
Shih-Jen Tsai,
Taipei Veterans General Hospital, Taiwan

***Correspondence:**
Yuichi Yamashita
yamay@ncnp.go.jp

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *13 January 2021*
Accepted: *07 June 2021*
Published: *07 July 2021*

Citation:
Yamaguchi H, Hashimoto Y, Sugihara G, Miyata J, Murai T, Takahashi H, Honda M, Hishimoto A and Yamashita Y (2021) Three-Dimensional Convolutional Autoencoder Extracts Features of Structural Brain Images With a "Diagnostic Label-Free" Approach: Application to Schizophrenia Datasets. Front. Neurosci. 15:652987. doi: 10.3389/fnins.2021.652987

INTRODUCTION

Deep learning (DL) has dramatically improved technology in speech recognition, image recognition, and many other fields (LeCun et al., 2015). Medical imaging can benefit greatly from recent progress in image classification and object detection using this cutting-edge technology (Esteva et al., 2019). In particular, as the global burden of psychiatric disorders increases (Olesen et al., 2012; Whiteford et al., 2013), psychiatric brain imaging studies using DL are anticipated to bring many benefits to society (Vieira et al., 2017). There are two major concerns about applying DL to psychiatric brain imaging: (1) treatment of the high dimensionality of data, and (2) the heterogeneity of psychiatric disorders (Feczko et al., 2019).

The dimensionality of raw magnetic resonance imaging (MRI) data is very high (often running into the millions), and large computer resources are required to analyze them. To reduce computational demands, in most neuroimaging studies, several feature extraction methods have been used. Region of interest (ROIs), one of the most popular feature extraction methods, has contributed to detecting various structural and functional abnormalities in the brains of patients with psychiatric disorders (Fornito et al., 2012; Fusar-Poli et al., 2012; Linden, 2012; Ratnanather et al., 2013). ROIs (often dozens or hundreds) are usually set based on neuroscience knowledge (Tzourio-Mazoyer et al., 2002). For example, average gray matter volumes or cortical thicknesses at specific ROIs are extracted as feature, and then the relationship between the feature and disease clinical information is analyzed (Desikan et al., 2006; Poldrack, 2007; Nelson et al., 2017). Even in the studies using DL, ROI-based features are often used as input (Vieira et al., 2017; Heinsfeld et al., 2018; Pinaya et al., 2019). In addition, many DL studies avoid using three-dimensional (3D) images directly, but instead, DL networks are trained using two-dimensional slices (Sarraf et al., 2017; Vieira et al., 2017; Aghdam et al., 2019). A limitation of these studies is that they ignore the 3D spatial information contained within the original MRI scans.

In recent years, with improvements in computer performance and refinement of computational techniques, studies have investigated how to treat 3D MRI scans as inputs to DL. For example, Wang et al. (2018) successfully discriminated Alzheimer's dementia from healthy subjects using 3D MRI data as input to DL. Similar attempts have been made for discriminating psychiatric disorders, including schizophrenia (Qureshi et al., 2019) and developmental disorders (Wang et al., 2019). Although these studies demonstrated that DL could apply to the analysis of 3D MRI data, discrimination-based approaches may be challenging due to the heterogeneity of psychiatric disorders.

Heterogeneity is one of the main challenges that current psychiatric research faces (Feczko et al., 2019). The current symptom-based definitions of psychiatric disorders, standardized in the Diagnostic and Statistical Manual of the American Psychiatric Association (DSM) (American Psychiatric Association., 2013) and the International Classification of Diseases (ICD) (World Health Organization., 1992), have been highlighted as lacking predictive and clinical validity due to genetic and clinical heterogeneity (Owen, 2014). For example, in schizophrenia, a recent study found evidence for significant overlapping of the relatively common risk variants tagged in genome-wide association studies (GWAS) between several psychiatric disorders, and there may also be lower genetic correlation within disorders (Lee et al., 2014). In addition, even in patients given the same diagnosis of schizophrenia, the severity of symptoms, response to medication, and prognosis often vary widely among patients (van Os and Kapur, 2009; Owen et al., 2016). Therefore, in psychiatric disorders research, a simple competition for discrimination accuracy based on the current disorder categories may be insufficient to elucidate on pathophysiology, although most current studies using DL are attempting to discriminate disease in healthy subjects (Plis et al., 2014; Vieira et al., 2017; Gao et al., 2021; Quaak et al., 2021).

One possible alternative direction for using DL techniques in psychiatric neuroimaging studies may be diagnostic label-free feature extraction. In the current study, we focus on an autoencoder (AE) as a DL algorithm that allows feature extraction without labels (Hinton, 2006). AE is supervised learning in a deep neural network having an output layer with the same data as the input layer. Since the input is as supervision, no labels are needed, unlike in general supervised learning.

Indeed, there are some studies that have used AE-based feature extraction for psychiatric neuroimaging. For example, Pinaya et al. (2019) extracted features from structural MRI scans using AE, i.e., without using diagnostic labels. The authors successfully predicted the age and gender of participants, and discriminated patients with autism spectrum disorders (ASD) and schizophrenia from healthy subjects. However, these studies used ROI-based features such as cortical thickness and functional connectivity as inputs to the AE. As such, the use of 3D brain images for inputs to the AE remains challenging, with a few exceptions. For example, Martinez-Murcia et al. (2020) extracted features from 3D brain MRI data of patients with Alzheimer's dementia using a 3D convolutional autoencoder (3D-CAE). They demonstrated that the extracted feature was useful for predicting age and Mini-Mental State Examination (MMSE) scores. This supports the efficacy of labeling free features based on 3D-CAE with MRI. However, particularly when investigating psychiatric disorders, the appropriate architecture of 3D-CAE has not been fully investigated.

The purpose of this study was to investigate an efficient 3D-CAE-based feature extraction for the neuroimaging of psychiatric disorders. More specifically, in the current study, we used datasets that included patients with schizophrenia, which has frequently been reported to be heterogeneous in previous neuroimaging studies (Sugihara et al., 2017). The key points of our study are: (1) to use 3D MRI data while preserving spatial information, and (2) diagnostic label-free feature extraction using 3D-CAE. For this purpose, we explored appropriate network structures of 3D-CAE by developing models with different network structures and comparing the predictive performance of clinical information by these extracted features.

MATERIALS AND METHODS

Experimental Overview

Figure 1 illustrates an experimental overview of our study. We used two datasets, including participants diagnosed with schizophrenia as well as healthy subjects: a dataset collected at Kyoto University (Kyoto dataset) and a public dataset, The Center for Biomedical Research Excellence (COBRE[1]) dataset. (1) Gray matter was first extracted from the structural MRI data as preprocessing. (2) We then trained 3D-CAE to extract a latent feature representation from structural MRI using the Kyoto dataset. Sixteen 3D-CAEs with varying network structures were prepared for investigation of the optimal network depth and complexity. (3) Subsequently, the COBRE dataset was used to evaluate the applicability to another dataset. (4) Finally, we evaluated whether the extracted feature retained clinical information by linear regression of the clinical information using the COBRE dataset.

Convolutional Autoencoder Training

An autoencoder is a kind of DL consisting of the encoder and the decoder. The encoder learns latent representations and reduces the dimension of the input. The decoder learns to reproduce the input as close as possible to the original using the latent representations. 3D-CAE extends this architecture by using convolutional layers that can extract features directly from 3D images (Guo et al., 2017; Nishio et al., 2017; Oh et al., 2019). The CAE has two main hyper parameters: the number of convolutional layers and the number of channels, which are the target of the current study.

The convolutional layers apply a filter to input to create feature maps that summarize the feature detected in the input. The feature maps are created for the number of channels. Since the convolutional layer generates feature maps while capturing the spatial information of the matrix, convolutional neural networks are beneficial to learning features of images. As the number of channels increases, the complexity of a model increases, but the number of dimensions of latent feature increase and requires a huge amount of computational power. Also, as the number of convolutions increases, the effective receptive field increases, thus allowing global and abstract feature to be extracted. The effective receptive field is a region of the original image that can potentially influence the activation of neurons (Le and Borji, 2017; Luo et al., 2017). If the effective receptive field is small, the feature will contain only local information of the brain, and if it is large, it will contain information on the whole brain.

[1] http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html

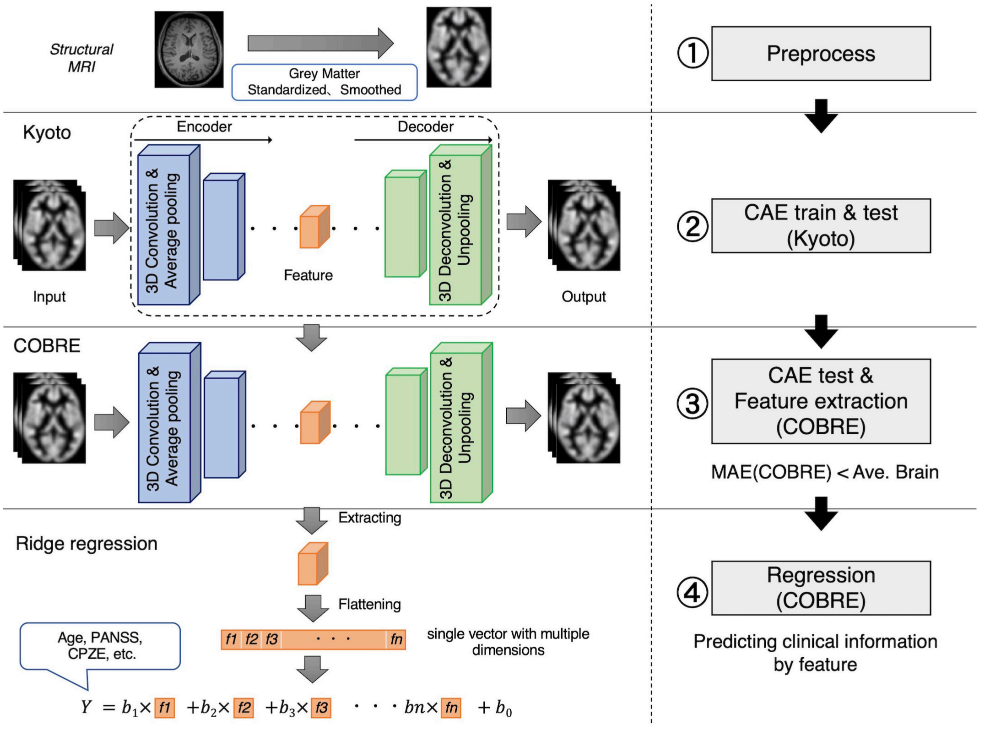

Structural MRI
Grey Matter Standardized、Smoothed
Kyoto
Encoder
Decoder
Input
3D Convolution & Average pooling
Feature
3D Deconvolution & Unpooling
Output
COBRE
3D Convolution & Average pooling
3D Deconvolution & Unpooling
Extracting
Ridge regression
Flattening
Age, PANSS, CPZE, etc.
f1 f2 f3 · · · fn
single vector with multiple dimensions
$Y = b_1 \times f1 + b_2 \times f2 + b_3 \times f3 \cdots bn \times fn + b_0$
① Preprocess
② CAE train & test (Kyoto)
③ CAE test & Feature extraction (COBRE)
MAE(COBRE) < Ave. Brain
④ Regression (COBRE)
Predicting clinical information by feature

FIGURE 1 | Experimental overview. 1. Preprocessing: The gray matter was extracted from the structural MRI and standardized and smoothed using SPM. 2. CAE training: A schematic diagram is shown. 3D images of the Kyoto dataset were input, feature was extracted, and the original image was reproduced. 3. Feature extraction: the model trained using the Kyoto dataset was adopted to the COBRE dataset without updating the weights. 4. Linear regression: Feature are extracted and flattened. Each extracted feature vectors were an explanatory variable, and demographic and clinical information were objective variables. Regression errors were evaluated to investigate whether the extracted features retain the information to predict demographic and clinical information. 3D, three-dimensional; CAE, convolutional autoencoder; COBRE, Center for Biomedical Research Excellence; CPZE, dose of antipsychotic medication; MRI, magnetic resonance imaging; SPM, Statistical Parametric Mapping.

In this study, these two hyperparameters were explored to investigate whether the total dimensions of the extracted feature and the size of the effective receptive field affected the relation of the feature to clinical information. As shown in **Figure 2**, the set of two convolution/deconvolution layers, and one pooling/unpooling layer was defined as a convolution/deconvolution "block." In this experiment, the number of blocks was set, ranging from 1 block to 4 blocks. In 4 blocks, the effective receptive field is the whole brain; in 3 blocks, it is about 30% of the brain (multiple lobes), in 2 blocks, it is 5% of the brain (multiple regions), and in 1 block it is 0.1% of the brain (1 region). The number of channels in the extraction layer was varied with 1, 4, 16, and 32 channels, but the number of channels for other layers were fixed at 32. The number of channels was considered limited to 32 due to the limitation of the current experiment's computational power. As a result, we created sixteen 3D-CAE models (4 block conditions × 4 channel conditions) to explore the effective range of hyperparameters for psychiatric brain imaging.

Other hyperparameters were fixed and common among models. The encoder was composed of convolution layers (a kernel size of 3 × 3 × 3 and a stride of 1) with rectified linear unit (ReLU) activations and average pooling layers (a kernel size of 2 × 2 × 2 and a stride of 2). The decoder was composed of convolution layers (a kernel size of 3 × 3 × 3 and a stride of 1) with ReLU activations and unpooling layers (a kernel size of 2 × 2 × 2 and a stride of 2). The loss function, consisting of the mean absolute error (MAE) between the input images and the reproduced images, was defined as follows:

$$Loss = \frac{1}{n} \sum \left| X_{input} - X_{reconstructed} \right| \tag{1}$$

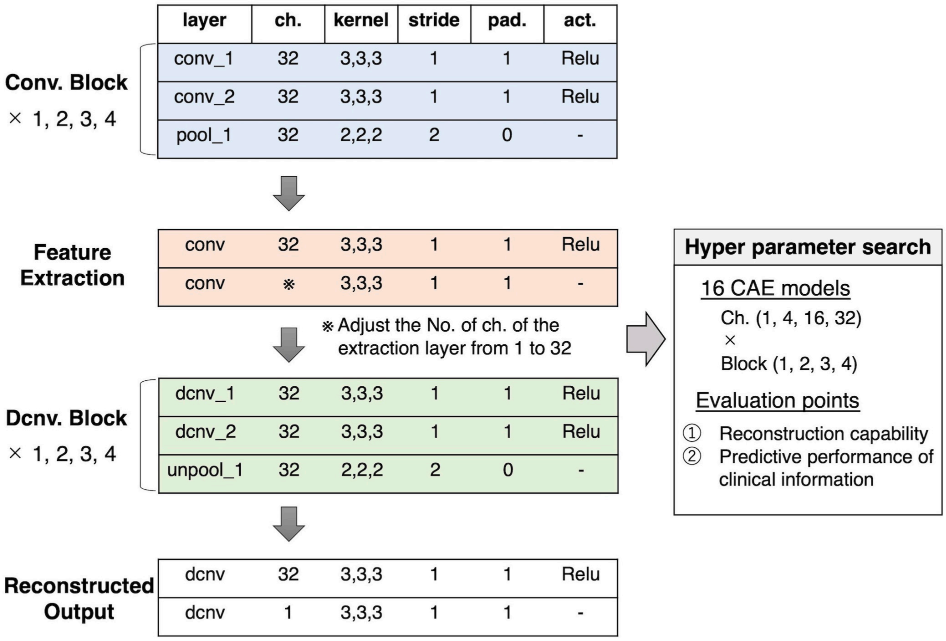
As an optimizer, we used a gradient-based method with adaptative learning rates called Adam (Kingma and Ba, 2015) (alpha = 0.0001, beta1 = 0.9, beta2 = 0.999) using mini-batches with a size of eight samples. The training process was performed with a maximum of 50,000 training iterations. We conducted the experiments in Python 3.6[2] using the Chainer v.5.4.0 library (Tokui et al., 2015).

We used a reference of training performances of 3D-CAEs, referred to as the "average brain," with which the model was assumed to output the average intensities of the training dataset regardless of the inputs. The average brain is one of the most trivial solutions where the network outputs an image without learning any information about individual differences of the inputs. The average brain was used as a reference point to indicate that the model at least reproduced individual differences. The signal intensities of voxel *i* of the average brain was determined as follows:

$$x_{ave\ i} = \frac{\sum_{s=0}^{n} x_{s,i}}{n} \tag{2}$$

where *s* is a sample from the training dataset and *n* is the number of samples.

[2] https://www.python.org/



Conv. Block
× 1, 2, 3, 4

layer	ch.	kernel	stride	pad.	act.
conv_1	32	3,3,3	1	1	Relu
conv_2	32	3,3,3	1	1	Relu
pool_1	32	2,2,2	2	0	-

Feature Extraction

conv	32	3,3,3	1	1	Relu
conv	※	3,3,3	1	1	-

※ Adjust the No. of ch. of the extraction layer from 1 to 32

Dcnv. Block
× 1, 2, 3, 4

dcnv_1	32	3,3,3	1	1	Relu
dcnv_2	32	3,3,3	1	1	Relu
unpool_1	32	2,2,2	2	0	-

Reconstructed Output

dcnv	32	3,3,3	1	1	Relu
dcnv	1	3,3,3	1	1	-

Hyper parameter search

16 CAE models
Ch. (1, 4, 16, 32)
×
Block (1, 2, 3, 4)

Evaluation points
① Reconstruction capability
② Predictive performance of clinical information

FIGURE 2 | Our proposed 3D-CAE architecture. One convolution/deconvolution block was defined as repeating two convolution/deconvolution layers and one pooling/unpooling layer. The number of blocks was set from 1 to 4. The number of channels in the extraction layer was set from 1 to 32. Sixteen patterns of models with different numbers of blocks and channels were developed. In order to explore the effective number of channels and blocks, the reproduction capability and relation to clinical information were evaluated. act., activation function; 3D-CAE, three-dimensional convolutional autoencoder; ch., channel; Conv., convolution; Dcnv, deconvolution; pad., padding; pool, pooling; Relu, Rectified Linear Unit; unpool, unpooling.

Regression Analysis With Demographic and Clinical Information

Using trained 3D-CAE, latent feature vector could be extracted, and then the feature vector was flattened. The number of dimensions of that feature vector ranged from millions to hundreds, depending on the model. The relationship between the extracted feature and the clinical information was examined using regression analysis, based on the assumption that if the extracted feature is "informative," it could help predict schizophrenia patients' clinical information. Therefore, we confirmed this by comparing the prediction performance of 3D-CAE-based features and conventional ROI-based features. The linear regression analysis was performed with clinical and demographic information as the objective variables and the feature vectors as the explanatory variables (see the lower part of **Figure 1**). Demographic and clinical information included age, scores of positive and negative symptoms (PANSS), the dose of antipsychotic medications [chlorpromazine equivalent (CPZE)], Wechsler Adult Intelligence Scale (WAIS), duration of illness, age at onset, and diagnosis. For the regression analysis, in order to reduce the effects of correlated variables we adopted ridge regression, one of regularized linear regression methods. In the regression analysis, we executed a fivefold cross-validation process whereby the COBRE dataset was randomly divided into five groups of samples (folds), and then samples from fourfolds were used for training the regression model, and the other fold was used for the test of the regression model. The fivefold cross-validation was repeated ten times. The performance of the regression model was evaluated using the root mean square error (RMSE). The diagnosis was evaluated using accuracy.

Differences in the performances of regression models were evaluated using the two-way (number of channels × number of blocks) analysis of variance (ANOVA). Subsequently, Tukey's multiple comparison test was performed for each group as a *post hoc* analysis. The level of significance was set to 0.05.

The 3D-CAE models were also compared with the ROI method. In the ROI method, using the automated anatomical labeling (AAL) template (Tzourio-Mazoyer et al., 2002), the GM was divided into 116 ROIs. The average intensities of each ROI were used as the ROI-based feature for regression analysis. The Student's *t*-test was performed to compare the proposed 3D-CAE model with the ROI method. The level of significance was set to 0.05.

By calculating the gradient of the neural network at the input T1-weighted image for each subject, it is possible to visualize which regions of the input have higher weights. In this study, we attempted to visualize the regions that contribute to predicting clinical information by calculating the gradient of a composite function of feature extraction and clinical information regression functions. The calculation of a saliency map for input image *x*, *M(x)*, was defined as follows.

$$M(x) = \partial R(S(x))/\partial x \tag{3}$$

Where, *S()* was a feature extraction function based on the 3D-CAE, and *R()* was a function predicting clinical information using linear regression. To refine the visualization, the gradients' calculation was repeated by adding Gaussian noise to the original image, similar to the technique used in SmoothGrad (Smilkov et al., 2017). The maps were then averaged by overall samples and divided by the standard deviation to obtain a *t*-value, and the values were finally converted to absolute values to yield a 3D saliency map.

Kyoto Dataset Description

A total of 172 subjects were investigated in this study, including 82 patients with schizophrenia and 90 healthy subjects. Patients were recruited from hospitals in Kyoto, Japan, and diagnosed by psychiatrists using the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV) (American Psychiatric Association., 1994) criteria for schizophrenia, confirmed with the patient edition of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID) (First et al., 1997). No patients had any comorbid DSM-IV Axis I disorder. The clinical symptoms of all patients were estimated using the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987). Healthy subjects were screened with the non-patient edition of the SCID, confirming no history of psychiatric disorders. Exclusion criteria for all individuals included a history of head trauma, neurological illness, serious medical or surgical illness, or substance abuse. Note that participants were already diagnosed in order to expedite the data collection, but the diagnostic labels were not used to train the networks.

All participants were scanned with a 3.0-Tesla Siemens Trio scanner (Siemens Healthineers, Erlangen, Germany). The scanning parameters of the T1-weighted 3D magnetization-prepared rapid gradient-echo (3D-MPRAGE) sequences were as follows: echo time (TE) = 4.38 ms; repetition time (TR) = 2,000 ms; inversion time (TI) = 990 ms; field of view (FOV) = 225 mm × 240 mm; acquisition matrix size = 240 × 256 × 208; resolution = 0.9375 × 0.9375 × 1.0 mm^3.

COBRE Dataset Description

In this study, the COBRE dataset, which is a public dataset, was acquired as a dataset with different scanning sites and parameters to the Kyoto University dataset. All the subjects were diagnosed and screened with the SCID. The clinical symptoms of all patients were estimated using the PANSS. Exclusion criteria for individuals included a history of head trauma, neurological illness, serious medical or surgical illness, or substance abuse. We included a total of 142 subjects from this database in our study, including 71 patients with schizophrenia and 71 healthy subjects.

MRI data were acquired using a 3.0-Tesla Siemens Tim Trio scanner (Siemens Healthineers, Erlangen, Germany). The scanning parameters of the T1-weighted 3D-MPRAGE sequences were as follows: TE = 1.64 ms; TR = 2,530 ms; TI = 900 ms; FOV = 256 mm × 256 mm; acquisition matrix size = 256 × 256 × 176; resolution = 1.0 × 1.0 × 1.0 mm^3.

Demographic and clinical characteristics of Kyoto and COBRE datasets are provided in **Supplementary Table 1**. There was no significant difference between the two datasets with the exception of the sex ratio.

Division of Train, Validation, and Test

The 3D-CAE was trained using the Kyoto dataset. The dataset was randomly partitioned into training data, validation data, and test data (138 subjects, 16 subjects, and 18 subjects, respectively). Training data, validation data, and test data were used for the training of the 3D-CAE, the validation of the model during training, and the final evaluation of generalizability within the datasets independent of the training and validation data, respectively. The COBRE dataset (142 subjects) was also used to evaluate the applicability of the network to another dataset.

The regression analysis was carried out using the COBRE dataset. The bias between MRI scanning sites might have affected the distribution of features extracted by 3D-CAE; thus, affecting the prediction error of the regression. Therefore, to avoid the scanning site effect, we used a single dataset for the regression. Then the fivefold cross-validation technique was applied. Namely, the COBRE dataset samples (142 subjects) were randomly divided into five subgroups (four groups for training and one group for validation) and cross-validated by changing the combinations of groups. This fivefold cross-validation process was repeated ten times. Note that only patients with schizophrenia had clinical information available for analysis, and regressions based on the clinical information were performed using data from patients with schizophrenia (71 subjects). The details for the division of data are shown in **Table 1**.

MRI Preprocessing

The preprocessing was conducted using Statistical Parametric Mapping (SPM12, Wellcome Department of Cognitive Neurology, London, United Kingdom[3]) with the Diffeomorphic Anatomical Registration Exponentiated Lie Algebra (DARTEL) registration algorithm (Ashburner, 2007). All of the T1 whole-brain structural MRI scans were segmented into gray matter (GM), white matter, and cerebrospinal fluid. Individual GM images were normalized to the standard Montreal Neurological Institute (MNI) template with a 1.5 × 1.5 × 1.5 mm^3 voxel size and modulated for GM volumes. All normalized GM images were smoothed with a Gaussian kernel of 8 mm full width at half maximum (FWHM). Subsequently, each image was cropped to remove the background as much as possible. The GM area was extracted from original images using a binary mask, created using SPM12. As a result, the size of input images to the 3D-CAE was 121 × 145 × 121 voxels.

Subsequently, the range of signal intensities in each image was normalized with a mean of 0 and a standard deviation of 1. The standardized value of voxel i in the sample s, $x'_{s,i}$, was calculated as follows:

$$x'_{s,i} = \begin{cases} \frac{x_{s,i} - \mu_s}{\sigma_s} & (i \in GM) \\ 0 & (otherwise) \end{cases} \tag{4}$$

where $x_{s,\ i}$ is the original value of intensity. μ_s and σ_s were average and standard deviation of all voxels contained in the GM area of sample s, respectively.

RESULTS

Technical Evaluations: Reproduction Capability Performance

Figure 3A shows a representative example of learning curves for the 3D-CAE with 16 channels and 3 blocks. Progressive decreases were shown not only with "train loss" (red line), but also "validation loss" (orange line) and "test loss" (green line); this indicated that the 3D-CAE successfully learned without overfitting. The level of MAEs were well below the level of the "average brain" (dashed line) (see section "Materials And Methods" for details). In addition, the curve for "COBRE loss" (blue line) showed a similar trend. This indicated that the 3D-CAE could be applied to MRI data from another site with different scanning parameters. Similar trends of learning curves were observed for the other fifteen 3D-CAEs with different hyperparameter settings.

Figure 3B summarized the reproduction performances (MAEs for the COBRE dataset) of the sixteen 3D-CAE models with respect to the number of channels and number of blocks. Regarding the number of blocks, it can be seen that the larger the number of blocks, the larger the reproduction error. This result is intuitively understandable, in that models with smaller blocks are easier to reconstruct because extracted latent features do not abstract the original image as much (**Figure 4**). Regarding the number of channels, although the differences were small, there was a tendency for the larger number of channels to be associated with smaller reproduction errors (see **Supplementary Table 2** for more details). This result is consistent with the fact that the models with more channels have more expressive capability.

Clinical Evaluation: Relation to Clinical Information

The efficacy of the proposed method was evaluated using linear regressions for predicting demographic and clinical information related to a psychiatric disorder, i.e., schizophrenia. Demographic and clinical information, including age, the dose of antipsychotic medication (CPZE), and scores of positive and negative symptoms (PANSS), were used as an objective variable, and all extracted features of 3D-CAE were used as explanatory variables. Feature using the ROI-based method was

TABLE 1 | Division of dataset.

		Kyoto	COBRE
3D-CAE			
(recon. error)	Train	✓	
	Validation	✓	
	Test	✓	✓
Regression			
(pred. error)	Train		✓
	Validation		✓

The Kyoto dataset was used to develop the 3D-CAE model and was divided into train, validation and test dataset. The COBRE dataset was prepared for regression. At regression, fivefold cross-validation was performed.

3D-CAE, three-dimensional convolutional autoencoder; COBRE, Center for Biomedical Research Excellence.

[3] https://www.fil.ion.ucl.ac.uk/spm/software/spm12/

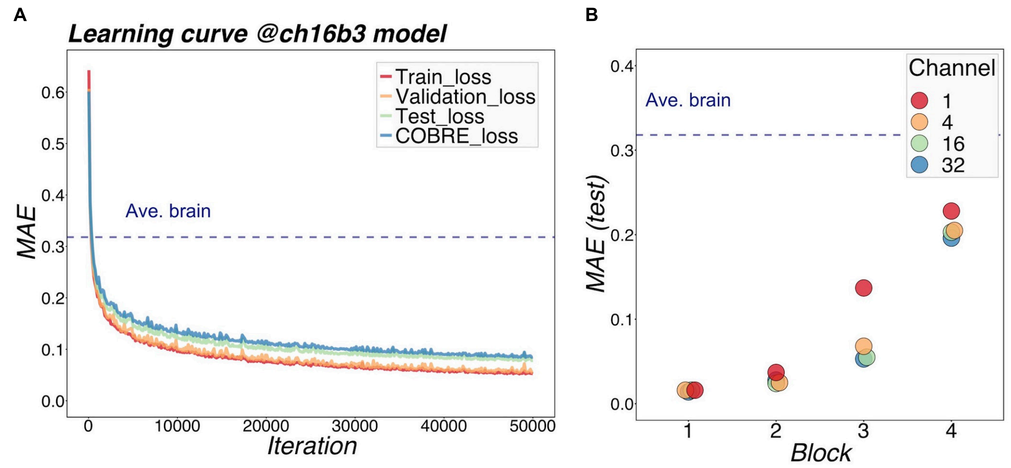


FIGURE 3 | Learning performance of models. **(A)** shows the learning loss curve for a 16-channel and 3-block model. The red line shows the training loss, indicating that the learning has progressed, and the loss has fallen sufficiently. The validation loss and test loss were also decreased, so the model was not overfitting. The blue line indicates the loss at the other site (COBRE), and the loss degraded as well. It can be seen that the MAE of our proposed models was well below the level of Ave.brain at which the model was assumed to output the average brain. This suggested that our 3D-CAE models have successfully reproduced the brain images with individual characteristics. Similar learning curves were found for other models. In **(B)**, the reproduction performance of each of the 16 models were compared. The relationships between MAE, number of channels, and number of blocks are shown. The horizontal axis indicates the number of blocks, which is color-coded by the number of channels. As the number of blocks increased, the MAE tended to be larger, and as the number of channels increased, the MAE tended to be slightly smaller. 3D-CAE, three-dimensional convolutional autoencoder; COBRE, Center for Biomedical Research Excellence; MAE, mean absolute error.

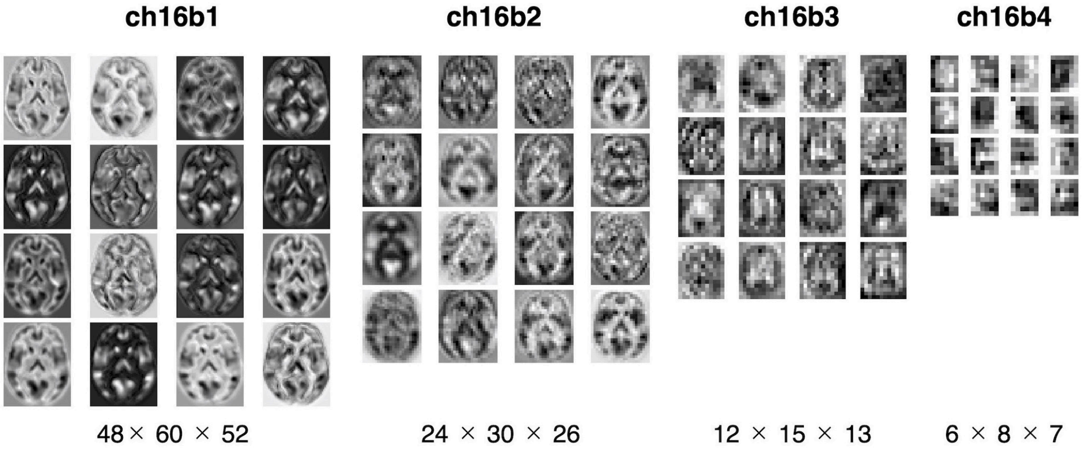


FIGURE 4 | Visualization of extracted feature. The extracted features were mapped for four models with 16 channels. From left to right: the model with one, two, three, and four blocks. The middle slices of the horizontal slice from 3D feature are shown. In the one-block model, the morphology of the brain can be seen, but with four blocks, the images are more abstract.

also used for comparison with the conventional method. A linear regression analysis was used as the simplest method to confirm if extracted features from 3D-CAEs with different hyperparameters (numbers of blocks and channels) preserved useful information. Each of the 16 3D-CAE models were analyzed 10 times, and the difference in predictive performance of the models was examined statistically.

Figure 5 illustrates a representative example of the regression analysis results. Differences in the performance of regression models (RMSE) with respect to the number of channels with 3 blocks (**Figures 5A–D**) and respect to the number of blocks with 16 channels (**Figures 5E–H**) were demonstrated as representative examples. The results of the comparison with the ROI method are shown in **Table 2**. The detailed results are described in **Supplementary Tables 3–5**, respectively.

Regarding the prediction of age, there were tendencies for the RMSEs to be smaller with increases in the number of channels (**Figure 5A**) and with decreasing number of blocks (**Figure 5B**). Indeed, statistical analysis revealed that there were significant differences between the models (channel: $p < 0.001$; block: $p < 0.001$). However, even the model with 32 channels and 1 block, which is considered one of the most predictive models, is equivalent to the ROI method ($p = 0.346$; **Table 2**), suggesting that for the

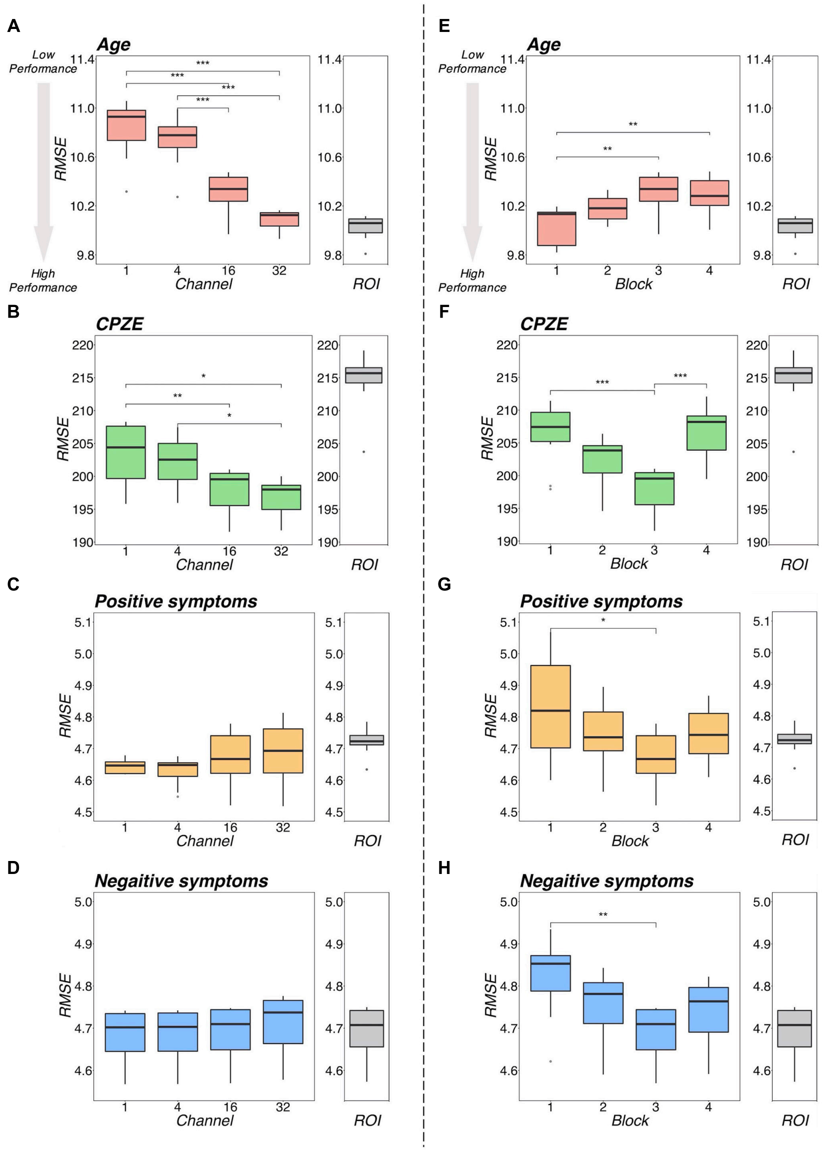


FIGURE 5 | Regression performance plot. The left side **(A–D)** shows the model differences by number of channels for the four models with 3 blocks as an example. The right side **(E–H)** shows the model differences by number of blocks for the four models, with 16 channels as representative examples. Regarding age, as shown in **(A,E)**, the RMSEs were smaller with increasing number of channels and decreasing number of blocks. Regarding CPZE, as shown in **(B,F)**, the RMSEs were smaller with increasing number of channels. On the other hand, the RMSEs may be smaller in block 3. Regarding positive symptoms and negative symptoms, as shown in **(C,D)**, there was no apparent trend in the number of channels. As shown in **(G,H)**, the RMSE may be smaller in block 3. The results of each regression with the ROI method is also included for reference. It suggests that a model with 3 blocks may be appropriate for extracting schizophrenia-related information.
***$p < 0.001$, **$p < 0.01$, *$p < 0.05$ (two-way analysis of variance followed by Tukey's multiple comparison test). CPZE, chlorpromazine equivalent; RMSE, root mean square error; ROI, region of interest.

prediction of age, 3D-CAE-based features were comparable to a conventional method.

In addition, the superiority of the 1 block condition was observed in the prediction of VIQ, PIQ and duration of illness (**Supplementary Tables 2–4**). However, 3D-CAEs with 1 block were not superior to the ROI method in predicting those information (VIQ: $p < 0.001$; PIQ: $p < 0.001$; duration of illness: $p = 0.100$; **Table 2**).

Regarding the prediction for CPZE, there was a tendency for the RMSEs to be smaller with increases in the number of channels (**Figure 5C**); on the other hand, the RMSEs were smallest with the condition of 3 blocks (**Figure 5D**). Statistical analysis revealed that there were significant differences between the models (channel: $p < 0.001$; block: $p < 0.001$). *Post hoc* analysis revealed that there were significant differences between 1 block and 3 blocks, and 3 blocks and 4 blocks. Moreover, the lowest level of RMSE of 3D-CAE was significantly lower than the RMSE from the ROI-based feature ($p < 0.001$; **Table 2**), indicating that for the prediction of CPZE, 3D-CAE based features outperformed a conventional method.

TABLE 2 | The results of the *t*-test.

16 channels and 3 blocks model

		3D-CAE (ch16b3)	**ROI**	***P*-value**
SZ-related clinical information	*CPZE*	197.85 (3.76)	214.75 (4.33)	$< 0.001^{***}$
	Positive symptoms	4.67 (0.09)	4.72 (0.04)	$0.088^{\dagger}$
	Negative symptoms	4.67 (0.07)	4.69 (0.07)	0.968
	Duration of illness	11.87 (0.10)	11.23 (0.16)	$< 0.001^{***}$
	Age of onset	7.00 (0.11)	7.47 (0.15)	$< 0.001^{***}$
	Diagnosis	0.668 (0.03)	0.634 (0.02)	0.005^{**}
Other information	*Age*	10.29 (0.18)	10.03 (0.10)	0.001^{**}
	VIQ	14.92 (0.17)	14.72 (0.05)	0.003^{**}
	PIQ	14.65 (0.11)	13.83 (0.09)	$< 0.001^{***}$

32 channels and 1 block model

		3D-CAE (ch32b1)	**ROI**	***P*-value**
SZ-related clinical information	*CPZE*	206.57 (4.61)	214.75 (4.33)	0.001^{**}
	Positive symptoms	4.84 (0.16)	4.72 (0.04)	0.037^{*}
	Negative symptoms	4.89 (0.10)	4.69 (0.07)	$< 0.001^{***}$
	Duration of illness	11.36 (0.17)	11.23 (0.16)	0.100
	Age of onset	7.05 (0.13)	7.47 (0.15)	$< 0.001^{***}$
	Diagnosis	0.632 (0.03)	0.634 (0.02)	0.868
Other information	*Age*	9.97 (0.16)	10.03 (0.10)	0.346
	VIQ	15.17 (0.17)	14.72 (0.05)	$< 0.001^{***}$
	PIQ	14.56 (0.13)	13.83 (0.09)	$< 0.001^{***}$

The differences between 3D-CAE and ROI are presented as mean (standard deviation) and p-value of RMSE. Regarding the diagnosis, it is presented as accuracy. The significantly better performances are marked in red. The 3D-CAE model with 16 channels and 3 blocks was superior to the ROI method in predicting CPZE, age of onset, and diagnosis. The model also appeared comparable or better than the ROI method in positive symptoms. The 3D-CAE model with 32 channels and 1 block was also superior to the ROI method in predicting the CPZE and age of onset. Meanwhile, that the model was comparable to the ROI method for age prediction is different from the model with 16 channels and 3 blocks. ${}^{***}p < 0.001$, ${}^{**}p < 0.01$, ${}^{*}p < 0.05$, ${}^{\dagger}p < 0.1$ (t-test).

3D-CAE, three-dimensional convolutional autoencoder; ROI, region of interest; SZ, Schizophrenia; CPZE, chlorpromazine equivalent.

Regarding the prediction of positive symptoms, there was no clear tendency with respect to the number of channels (**Figure 5E**). On the other hand, with respect to the number of blocks, the RMSEs seemed to be the smallest with the condition of 3 blocks (**Figure 5F**). Statistical analysis indicated that there were significant differences between the models (channel: $p < 0.001$; block: $p < 0.001$). *Post hoc* analysis revealed that there were significant differences between 1 block and 3 blocks. Similar trends could be observed in the prediction of negative symptoms (**Figures 5G,H**), where there were significant differences between the models (channel: $p < 0.001$; block: $p < 0.001$). In comparison to the conventional method, the 3D-CAE model with 3 blocks showed a trend toward a smaller prediction error for positive symptoms than the ROI method ($p = 0.088$; **Table 2**), the mean RMSE (SD) was 4.67 (0.09) and 4.72 (0.04), respectively, suggesting that the 3D-CAE might be comparable or better than the ROI method. Regarding the prediction of negative symptoms, there was no significant difference between 3D-CAE and the conventional method ($p = 0.968$; **Table 2**).

In addition, the superiority of the 3 blocks condition was observed in the prediction of age of onset and diagnosis (**Supplementary Tables 2–4**). Furthermore, 3D-CAEs with 3 blocks performed better than the ROI method in predicting those clinical information (age of onset: $p < 0.001$; diagnosis: $p = 0.005$; **Table 2**).

To summarize the regression analysis results, in terms of clinical information related to schizophrenia, specifically for predicting CPZE, positive symptom score, age of onset, and diagnosis, 3D-CAE with 3 blocks had better prediction than other numbers of blocks models, regardless of the number of channels. In addition, 3D-CAE with 3 blocks performed better than the ROI method in predicting clinical information. On the other hand, in terms of information not directly related to schizophrenia, such as age and intelligence, 3D-CAE with 1 block had better prediction than 3D-CAE with other numbers of blocks, regardless of the number of channels. However, 3D-CAE with 1 block did not perform better than the ROI method in predicting information not directly related to schizophrenia.

The saliency map was calculated to examine the correspondence between the features and the brain (**Figure 6**). The map showed that the regions contributing to CPZE prediction using 3D-CAE were the cerebellum, right middle temporal gyrus, the insula, posterior cingulate cortex, and precuneus. The regions that contributed to predicting the positive symptoms were found to the cerebellum, right inferior temporal gyrus, the insula, anterior and middle cingulate cortex. The other visualization results are described in **Supplementary Figure 1**.

DISCUSSION

We have shown that (1) the proposed 3D-CAEs successfully reproduced 3D MRI data with sufficiently low errors, and (2) the diagnostic label-free features extracted using 3D-CAE retained the relation of various clinical information. In addition, we explored the appropriate hyperparameter range of 3D-CAE, and our results suggest that a model with 3 blocks-based features might preserve information related to the medication dose and the severity of positive symptoms in patients with schizophrenia.

The reproduction errors of 3D-CAE were lower than the average brain level, indicating that the proposed 3D-CAEs successfully reproduced 3D brain MRI data with individual characteristics. In addition, the 3D-CAE trained with the Kyoto dataset was applicable to the COBRE dataset with different

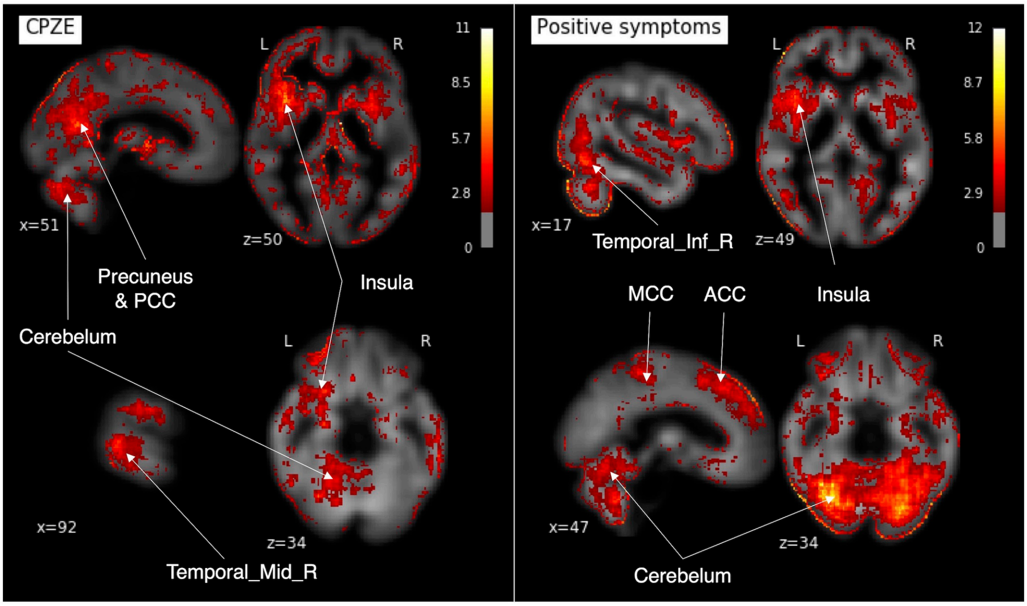


FIGURE 6 | The saliency maps. In our developed 3D-CAE model, the saliency maps of the signals contributing to the prediction of each clinical information were obtained by calculating the gradient of the neural network. 3D-CAE, three-dimensional convolutional autoencoder; PCC, posterior cingulate cortex; MCC, middle cingulate cortex; ACC, anterior cingulate cortex; Temporal_Mid_R, right middle temporal gyrus; Temporal_Inf_R, right inferior temporal gyrus.

scanners and scanning parameters. Although the current study was tested using only two datasets, the results suggested that the proposed method may have applicability to data from multiple sites and scanners, itself a challenging issue in neuroimaging studies (Jovicich et al., 2009; Schnack et al., 2010; Fortin et al., 2018; Dewey et al., 2019; Yamashita et al., 2019).

Regression analyses demonstrated that the 3D-CAE-based feature was comparable or more effective than the ROI-based feature in predicting the medication dose and the severity of positive symptoms in patients with schizophrenia, even though 3D-CAE-based features were extracted without using a diagnostic label of schizophrenia. Because this approach enabled us to extract neuroimaging features of individuals without information on the clinical diagnosis, it can be useful for heterogeneous population data. Furthermore, using this approach, we were able to predict clinical variables. These imply that our approach in this study could be an alternative method to conventional methods based on categorical diagnostic information. This study showed that the prediction of CPZE, positive symptoms, and age of onset might be more improved in 3D-CAE than ROI. These are clinically meaningful because the model would help clinicians decide the treatment plan by predicting based on an objective indicator. In contrast, the current medication dose is mainly adjusted based on the patient's self-reported condition.

Regarding the number of channels, 16— to 32-channel models demonstrated better performance. This is easy to understand because the more channels the model has, the more expressive it is (Zhu et al., 2019). However, since increasing number of channels inevitably results in increasing computational power needs, estimation of the appropriate number of channels is still important. Our results suggest that the number of channels may be sufficient at 16 or 32 for reproducing structural brain MRI scans. Regarding the number of blocks, our results indicated that information from a local receptive field (small number of block) was sufficient for predicting age. However, predicting schizophrenia-related clinical data required information from more global receptive fields (larger block numbers, such as 3-block). As the number of blocks increase, the effective receptive fields expand, and the global feature of the brain can be extracted (Szegedy et al., 2015; Le and Borji, 2017; Luo et al., 2017). In our model, the 3 blocks model contained eight convolutional layers, and effective receptive fields of the feature unit were about 68 × 68 × 68 voxels, corresponding to about 30% of the brain. This fact is consistent with the previous neuroimaging studies showing that the medication dose and symptoms severity are associated with the volume of multiple brain regions, including the temporal lobe, frontal lobe, and various subcortical regions (García-Martí et al., 2008; Palaniyappan et al., 2013; Van Erp et al., 2016; van Erp et al., 2018; Bullmore, 2019; Fan et al., 2019). The 3D-CAE-based feature's superiority may be related to the detection of local signal interactions inherent in the convolutional methods; this contrasts with the ROI-based method, in which signals within each ROI are averaged and the interactions of local signals are discarded.

In our model, the saliency map showed that the cerebellum, temporal lobe, cingulate gyrus, and insular cortex had greater contributions in predicting the severity of symptoms and dose of antipsychotic medication. The present study results were consistent with the results of previous studies showing that positive symptoms and CPZE correlated with cortical thickness thinning in the temporal lobe (van Erp et al., 2018), and that cerebellar atrophy was associated with positive symptoms (Cierpka et al., 2017). The insular and cingulate cortices, which were shown to be significant contributors to clinical variables in the present study, have been repeatedly reported to be reduced in the regional brain volume in schizophrenia (Glahn et al., 2008; Takayanagi et al., 2013; Gupta et al., 2015; Uwatoko et al., 2015). However, the relationship between these areas and positive symptoms and CPZE requires further investigation. As a side note, because the relatively high values of the edge of the brain may be influenced by the traits of Smoothgrad (Smilkov et al., 2017) that emphasize the edge, it was difficult to consider them from a neuroimaging study perspective.

There are some limitations to our study. First, this study only explored a limited range of hyperparameters. In CAE, there are several hyperparameters than those explored, such as activate function, optimizer, and learning rate. However, because we focused on the total dimension of the extracted features and the effective receptive field's size, the numbers of blocks and channels were explored as the target variables. In addition, the exploration range of hyperparameters was limited due to practical reasons including the computational power and costs of the experiments.

Second, the differences in preprocessing of neuroimaging data may affect the robustness of the study results. In this study, we employed the standard preprocessing methods (e.g., image resolution, standardization, smoothing), which have been used in neuroimaging studies, such as voxel-based morphometry (Ashburner and Friston, 2000). Nevertheless, further studies may evaluate the effects of the preprocessing methods on results.

Third, the datasets used in this study only included patients diagnosed with schizophrenia as well as healthy subjects. Considering the heterogeneity of psychiatric disorders, it will be necessary to examine the applicability of diagnostic label-free feature extraction using 3D-CAE to other psychiatric disorders in the future.

Fourth, regressions were used to predict clinical and demographic scores, but the 3D-CAE-based feature outperformed the feature of the ROI does not necessarily prove that the predictive value generated is clinically useful. In the present study, the main goal was feature extraction, and only simple regression was used for prediction. The additional experiments with the development of a fine-tuned model and evaluation using longitudinal data of disease process are needed in the future. These may improve clinical decisions for assessing patients' prognosis and estimating an appropriate medication dose.

In this paper, we presented 3D-CAE-based feature extraction for brain structural imaging of psychiatric disorders. We found that 3D-CAE can extract features that retained their relation to clinical information from 3D MRI data without diagnostic labels. Our data suggest that 3D-CAE models with effective hyperparameter settings may extract information related to the medication dose and the severity of symptoms in patients with schizophrenia. The feature extraction without using diagnostic labels based on the current diagnostic criteria is scientifically significant and may lead to the development of alternative data-driven diagnostic criteria.

DATA AVAILABILITY STATEMENT

All data generated or analyzed during this study are included in this published article. The primary data can be obtained from public databases, including the Decoded Neurofeedback (DecNef) Project Brain Data Repository (https://bicr-resource.atr.jp/srpbs1600/) and the Centers for Biomedical Research Excellence (COBRE; http://fcon_1000.projects.nitrc.org/indi/retro/cobre.html).

ETHICS STATEMENT

All study participants signed an informed consent form. The study was performed in accordance with the current Ethical Guidelines for Medical and Health Research Involving Human Subjects in Japan and was approved by the Committee on Medical Ethics of Kyoto University and National Center of Neurology and Psychiatry.

AUTHOR CONTRIBUTIONS

HY, YH, and YY conceived, designed the research, and drafted the manuscript. HY and YH conducted the deep learning experiments and analyzed the data. GS, JM, TM, and HT collected MRI data. GS, JM, TM, HT, MH, and AH provided critical revisions. All authors contributed to and have approved the final manuscript.

FUNDING

This work was partially supported by JSPS KAKENHI Grant Nos. JP17H00740, JP17H04248, JP17H06039, JP18KT0021, JP18K07597, JP19H04998, JP19K17077, JP20H00001, and JP20H00625, Grant-in-Aid for Scientific Research on Innovative Areas from the MEXT Grant No. JP20H05064, Strategic International Brain Science Research Promotion Program by the Japan Agency for Medical Research and Development Grant No. 19dm0307008h0002, JST CREST Grant No. JPMJCR16E2 and a Grant from SENSHIN Medical Research Foundation.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://www.biorxiv.org/content/10.1101/2020.08.24.213447v1. We thank all the authors of the included studies. We also would like to thank Editage (www.editage.com) for English language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.652987/full#supplementary-material

REFERENCES

Aghdam, M. A., Sharifi, A., and Pedram, M. M. (2019). Diagnosis of autism spectrum disorders in young children based on resting-state functional magnetic resonance imaging data using convolutional neural networks. *J. Digit. Imaging* 32, 899–918. doi: 10.1007/s10278-019-00196-1

American Psychiatric Association. (1994). *Diagnostic And Statistical Manual Of Mental Disorders?: DMS-IV*. Washington, DC: American Psychiatric Publishing.

American Psychiatric Association. (2013). *Diagnostic And Statistical Manual Of Mental Disorders (DSM-5)*. Washington, DC: American Psychiatric Publishing.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage* 38, 95–113. doi: 10.1016/j.neuroimage.2007.07.007

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry – the methods. *Neuroimage* 11, 805–821. doi: 10.1006/nimg.2000.0582

Bullmore, E. (2019). Cortical thickness and connectivity in schizophrenia. *Am. J. Psychiatry* 176, 505–506. doi: 10.1176/appi.ajp.2019.19050509

Cierpka, M., Wolf, N. D., Kubera, K. M., Schmitgen, M. M., Vasic, N., Frasch, K., et al. (2017). Cerebellar contributions to persistent auditory verbal hallucinations in patients with schizophrenia. *Cerebellum* 16, 964–972. doi: 10.1007/s12311-017-0874-5

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

Dewey, B. E., Zhao, C., Reinhold, J. C., Carass, A., Fitzgerald, K. C., Sotirchos, E. S., et al. (2019). DeepHarmony: a deep learning approach to contrast harmonization across scanner changes. *Magn. Reson. Imaging* 64, 160–170. doi: 10.1016/j.mri.2019.05.041

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z

Fan, F., Xiang, H., Tan, S., Yang, F., Fan, H., Guo, H., et al. (2019). Subcortical structures and cognitive dysfunction in first episode schizophrenia. *Psychiatry Res. Neuroimaging* 286, 69–75. doi: 10.1016/j.pscychresns.2019.01.003

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., and Fair, D. A. (2019). The Heterogeneity problem: approaches to identify psychiatric subtypes. *Trends Cogn. Sci.* 23, 584–601. doi: 10.1016/j.tics.2019.03.009

First, M. B., Spitzer, R. L., Gibbon, M., and Williams, J. B. W. (1997). *Structured Clinical Interview for DSM-IV Axis I Disorders SCID-I*. Washington, DC: American Psychiatric Publishing.

Fornito, A., Zalesky, A., Pantelis, C., and Bullmore, E. T. (2012). Schizophrenia, neuroimaging and connectomics. *Neuroimage* 62, 2296–2314. doi: 10.1016/j.neuroimage.2011.12.090

Fortin, J.-P., Cullen, N., Sheline, Y. I., Taylor, W. D., Aselcioglu, I., Cook, P. A., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* 167, 104–120. doi: 10.1016/j.neuroimage.2017.11.024

Fusar-Poli, P., Howes, O., Bechdolf, A., and Borgwardt, S. (2012). Mapping vulnerability to bipolar disorder: a systematic review and meta-analysis of neuroimaging studies. *J. Psychiatry Neurosci.* 37, 170–184. doi: 10.1503/jpn.110061

Gao, J., Chen, M., Li, Y., Gao, Y., Li, Y., Cai, S., et al. (2021). Multisite autism spectrum disorder classification using convolutional neural network classifier and individual morphological brain networks. *Front. Neurosci.* 14:629630. doi: 10.3389/fnins.2020.629630

García-Martí, G., Aguilar, E. J., Lull, J. J., Martí-Bonmatí, L., Escartí, M. J., Manjón, J. V., et al. (2008). Schizophrenia with auditory hallucinations: a voxel-based morphometry study. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 32, 72–80. doi: 10.1016/j.pnpbp.2007.07.014

Glahn, D. C., Laird, A. R., Ellison-Wright, I., Thelen, S. M., Robinson, J. L., Lancaster, J. L., et al. (2008). Meta-analysis of gray matter anomalies in schizophrenia: application of anatomic likelihood estimation and network analysis. *Biol. Psychiatry* 64, 774–781. doi: 10.1016/j.biopsych.2008.03.031

Guo, X., Liu, X., Zhu, E., and Yin, J. (2017). "Deep clustering with convolutional autoencoders," in *Neural Information Processing. ICONIP 2017. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, eds D. Liu, S. Xie, Y. Li, D. Zhao, and E. S. El-Alfy (Cham: Springer), 373–382. doi: 10.1007/978-3-319-70096-0_39

Gupta, C. N., Calhoun, V. D., Rachakonda, S., Chen, J., Patel, V., Liu, J., et al. (2015). Patterns of gray matter abnormalities in schizophrenia based on an international mega-analysis. *Schizophr. Bull.* 41, 1133–1142. doi: 10.1093/schbul/sbu177

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuroimage Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Hinton, G. E. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Jovicich, J., Czanner, S., Han, X., Salat, D., van der Kouwe, A., Quinn, B., et al. (2009). MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *Neuroimage* 46, 177–192. doi: 10.1016/j.neuroimage.2009.02.010

Kay, S. R., Fiszbein, A., and Opler, L. A. (1987). The positive and negative syndrome scale (PANSS) for schizophrenia. *Schizophr. Bull.* 13, 261–276. doi: 10.1093/schbul/13.2.261

Kingma, D. P., and Ba, J. L. (2015). "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015* (Ithaca, NY).

Le, H., and Borji, A. (2017). What are the Receptive, Effective Receptive, and Projective Fields of Neurons in Convolutional Neural Networks?. Available onlne at: http://arxiv.org/abs/1705.07049 (accessed May, 2020).

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

Lee, S., Ripke, S., Neale, B. M., Faraone, S., Purcell, S., Rh, P., et al. (2014). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs cross-disorder group of the psychiatric genomics consortium. *Nat. Genet.* 45, 984–994. doi: 10.1038/ng.2711

Linden, D. E. J. (2012). The challenges and promise of neuroimaging in psychiatry. *Neuron* 73, 8–22. doi: 10.1016/j.neuron.2011.12.014

Luo, W., Li, Y., Urtasun, R., and Zemel, R. (2017). Understanding the effective receptive field in deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 4905–4913. Available at: http://arxiv.org/abs/1701.04128 (accessed May, 2020).

Martinez-Murcia, F. J., Ortiz, A., Gorriz, J.-M., 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

Nelson, B. G., Bassett, D. S., Camchong, J., Bullmore, E. T., and Lim, K. O. (2017). Comparison of large-scale human brain functional and anatomical networks in schizophrenia. *Neuroimage Clin.* 15, 439–448. doi: 10.1016/j.nicl.2017.05.007

Nishio, M., Nagashima, C., Hirabayashi, S., Ohnishi, A., Sasaki, K., Sagawa, T., et al. (2017). Convolutional auto-encoders for image denoising of ultra-low-dose CT. *Heliyon* 3:e00393. doi: 10.1016/j.heliyon.2017.e00393

Oh, K., Kim, W., Shen, G., Piao, Y., Kang, N. I., Oh, I. S., et al. (2019). Classification of schizophrenia and normal controls using 3D convolutional neural network and outcome visualization. *Schizophr. Res.* 212, 186–195. doi: 10.1016/j.schres.2019.07.034

Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H.-U., and Jönsson, B. (2012). The economic cost of brain disorders in Europe. *Eur. J. Neurol.* 19, 155–162. doi: 10.1111/j.1468-1331.2011.03590.x

Owen, M. J. (2014). New approaches to psychiatric diagnostic classification. *Neuron* 84, 564–571. doi: 10.1016/j.neuron.2014.10.028

Owen, M. J., Sawa, A., and Mortensen, P. B. (2016). Schizophrenia. *Lancet* 388, 86–97. doi: 10.1016/S0140-6736(15)01121-6

Palaniyappan, L., Marques, T. R., Taylor, H., Handley, R., Mondelli, V., Bonaccorso, S., et al. (2013). Cortical folding defects as markers of poor treatment response in first-episode psychosis. *JAMA Psychiatry* 70, 1031–1040. doi: 10.1001/jamapsychiatry.2013.203

Pinaya, W. H. L., Mechelli, A., and Sato, J. R. (2019). Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: a large-scale multi-sample study. *Hum. Brain Mapp.* 40, 944–954. doi: 10.1002/hbm.24423

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

Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Soc. Cogn. Affect. Neurosci.* 2, 67–70. doi: 10.1093/scan/nsm006

Quaak, M., van de Mortel, L., Mani Thomas, R., and van Wingen, G. (2021). Deep learning applications for the classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis. *Neuroimage Clin.* 30:102584. doi: 10.1016/j.nicl.2021.102584

Qureshi, M. N. I., Oh, J., and Lee, B. (2019). 3D-CNN based discrimination of schizophrenia using resting-state fMRI. *Artif. Intell. Med.* 98, 10–17. doi: 10.1016/j.artmed.2019.06.003

Ratnanather, J. T., Poynton, C. B., Pisano, D. V., Crocker, B., Postell, E., Cebron, S., et al. (2013). Morphometry of superior temporal gyrus and planum temporale in schizophrenia and psychotic bipolar disorder. *Schizophr. Res.* 150, 476–483. doi: 10.1016/j.schres.2013.08.014

Sarraf, S., DeSouza, D. D., Anderson, J., and Tofighi, G. (2017). DeepAD: Alzheimer's disease classification via deep convolutional neural networks using MRI and fMRI. *bioRxiv*[Preprint] 070441. doi: 10.1101/070441

Schnack, H. G., van Haren, N. E. M., Brouwer, R. M., van Baal, G. C. M., Picchioni, M., Weisbrod, M., et al. (2010). Mapping reliability in multicenter MRI: voxel-based morphometry and cortical thickness. *Hum. Brain Mapp.* 31, 1967–1982. doi: 10.1002/hbm.20991

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. (2017). *SmoothGrad: Removing Noise by Adding Noise*. Available online at: http://arxiv.org/abs/1706.03825 (accessed June, 2019).

Sugihara, G., Oishi, N., Son, S., Kubota, M., Takahashi, H., and Murai, T. (2017). Distinct patterns of cerebral cortical thinning in schizophrenia: a neuroimaging data-driven approach. *Schizophr. Bull.* 43, 900–906. doi: 10.1093/schbul/sbw176

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington, DC: IEEE Computer Society), 1–9. doi: 10.1109/CVPR.2015.7298594

Takayanagi, M., Wentz, J., Takayanagi, Y., Schretlen, D. J., Ceyhan, E., Wang, L., et al. (2013). Reduced anterior cingulate gray matter volume and thickness in subjects with deficit schizophrenia. *Schizophr. Res.* 150, 484–490. doi: 10.1016/j.schres.2013.07.036

Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). "Chainer: a next-generation open source framework for deep learning," in *Proceedings of Workshop On Machine Learning Systems (LearningSys) in the 29th Annual Conference On Neural Information Processing Systems (NIPS)*, (San Diego, USA: The Neural Information Processing Systems Foundation), 1–6.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Uwatoko, T., Yoshizumi, M., Miyata, J., Ubukata, S., Fujiwara, H., Kawada, R., et al. (2015). Insular gray matter volume and objective quality of life in schizophrenia. *PLoS One* 10:e0142018. doi: 10.1371/journal.pone.0142018

Van Erp, T. G. M., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., et al. (2016). Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the ENIGMA consortium. *Mol. Psychiatry* 21, 547–553. doi: 10.1038/mp.2015.63

van Erp, T. G. M., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., et al. (2018). Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (ENIGMA) consortium. *Biol. Psychiatry* 84, 644–654. doi: 10.1016/j.biopsych.2018.04.023

van Os, J., and Kapur, S. (2009). Schizophrenia. *Lancet* 374, 635–645. doi: 10.1016/S0140-6736(09)60995-8

Vieira, S., Pinaya, W. H. L., 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

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

Wang, Z., Sun, Y., Shen, Q., and Cao, L. (2019). Dilated 3D convolutional neural networks for brain MRI data classification. *IEEE Access* 7, 134388–134398. doi: 10.1109/ACCESS.2019.2941912

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., et al. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *Lancet* 382, 1575–1586. doi: 10.1016/S0140-6736(13)61611-6

World Health Organization. (1992). *International Statistical Classification Of Diseases And Related Health Problems*. Geneva: World Health Organization.

Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., et al. (2019). Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLoS Biol* 17:e3000042. doi: 10.1371/journal.pbio.3000042

Zhu, H., An, Z., Yang, C., Hu, X., Xu, K., and Xu, Y. (2019). *Rethinking the Number of Channels for the Convolutional Neural Network*. Available online at: http://arxiv.org/abs/1909.01861 (accessed May, 2020).

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 | Frontiers in Neuroscience

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fnins.2022.889808

Uncertainty-Aware and Lesion-Specific Image Synthesis in Multiple Sclerosis Magnetic Resonance Imaging: A Multicentric Validation Study

Tom Finck[1†], Hongwei Li[2†], Sarah Schlaeger[1], Lioba Grundl[1], Nico Sollmann[1,3], Benjamin Bender[4], Eva Bürkle[4], Claus Zimmer[1], Jan Kirschke[1], Björn Menze[2], Mark Mühlau[5,6] and Benedikt Wiestler[1,2]*

[1] Department of Diagnostic and Interventional Neuroradiology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany, [2] Image-Based Biomedical Modeling, Technical University of Munich, Munich, Germany, [3] Department of Diagnostic and Interventional Radiology, University Hospital Ulm, Ulm, Germany, [4] Department of Diagnostic and Interventional Neuroradiology, Universitätsklinikum Tübingen, Tübingen, Germany, [5] TUM-Neuroimaging Center, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany, [6] Department of Neurology, School of Medicine, Klinikum rechts der Isar, Technical University of Munich, Munich, Germany

OPEN ACCESS

Edited by:
*Yogesh Rathi,
Harvard Medical School,
United States*

Reviewed by:
*Fan Zhang,
Harvard Medical School,
United States
Guoping Xu,
Wuhan Institute of Technology, China*

Correspondence:
*Tom Finck
tom.finck@tum.de*

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *04 March 2022*
Accepted: *04 April 2022*
Published: *26 April 2022*

Citation:
Finck T, Li H, Schlaeger S, Grundl L, Sollmann N, Bender B, Bürkle E, Zimmer C, Kirschke J, Menze B, Mühlau M and Wiestler B (2022) Uncertainty-Aware and Lesion-Specific Image Synthesis in Multiple Sclerosis Magnetic Resonance Imaging: A Multicentric Validation Study. Front. Neurosci. 16:889808. doi: 10.3389/fnins.2022.889808

Generative adversarial networks (GANs) can synthesize high-contrast MRI from lower-contrast input. Targeted translation of parenchymal lesions in multiple sclerosis (MS), as well as visualization of model confidence further augment their utility, provided that the GAN generalizes reliably across different scanners. We here investigate the generalizability of a refined GAN for synthesizing high-contrast double inversion recovery (DIR) images and propose the use of uncertainty maps to further enhance its clinical utility and trustworthiness. A GAN was trained to synthesize DIR from input fluid-attenuated inversion recovery (FLAIR) and T1w of 50 MS patients (training data). In another 50 patients (test data), two blinded readers (R1 and R2) independently quantified lesions in synthetic DIR (synthDIR), acquired DIR (trueDIR) and FLAIR. Of the 50 test patients, 20 were acquired on the same scanner as training data (internal data), while 30 were scanned at different scanners with heterogeneous field strengths and protocols (external data). Lesion-to-Background ratios (LBR) for MS-lesions vs. normal appearing white matter, as well as image quality parameters were calculated. Uncertainty maps were generated to visualize model confidence. Significantly more MS-specific lesions were found in synthDIR compared to FLAIR (R1: 26.7 ± 2.6 vs. 22.5 ± 2.2 $p < 0.0001$; R2: 22.8 ± 2.2 vs. 19.9 ± 2.0, $p = 0.0005$). While trueDIR remained superior to synthDIR in R1 [28.6 ± 2.9 vs. 26.7 ± 2.6 ($p = 0.0021$)], both sequences showed comparable lesion conspicuity in R2 [23.3 ± 2.4 vs. 22.8 ± 2.2 ($p = 0.98$)]. Importantly, improvements in lesion counts were similar in internal and external data. Measurements of LBR confirmed that lesion-focused GAN training significantly improved lesion conspicuity. The use of uncertainty maps furthermore helped discriminate between MS lesions and artifacts. In conclusion, this multicentric study confirms the external validity of a lesion-focused Deep-Learning tool aimed at MS imaging. When implemented, uncertainty maps are promising to increase the trustworthiness of synthetic MRI.

Keywords: magnetic resonance imaging, neuroradiology, multiple sclerosis, deep learning – artificial neural network (DL-ANN), double inversion recovery (DIR), synthetic MRI, artificial intelligence (AI)

INTRODUCTION

Magnetic resonance imaging (MRI) plays a central role in the management of patients with multiple sclerosis (MS), a neuroinflammatory disease with rising incidence that remains the most common cause of non-traumatic disability in the young (GBD 2016 Multiple Sclerosis Collaborators, 2019). MRI techniques have been developed to detect specific aspects of MS pathophysiology; double inversion recovery (DIR) imaging is exemplary of a sequence that improves lesion detection, in particular within the juxtacortical region. Through numerous studies, the superiority of DIR compared to established MRI sequences such as T2w or fluid-attenuated inversion recovery (FLAIR) sequences in depicting inflammatory white matter lesions has been validated (Geurts et al., 2005; Wattjes et al., 2007). Lengthy acquisition times and high technical requirements have, however, hindered the widespread use of DIR.

Recently, it has been shown that synthesizing DIR images with generative adversarial networks (GANs), a deep learning (DL) architecture with great potential for image synthesis, is feasible and improves lesion detection compared to FLAIR and T2w sequences (Finck et al., 2020; Bouman et al., 2021). Nonetheless, and in particular as MS lesions typically are small, GANs are at risk to synthesize images of high morphologic similarity to the target image, while failing to translate the clinically important MS lesions. Domain knowledge, i.e., the ability of a GAN to learn about the pathology-specific anomalies it should map, might open the door for further customization and improvements in this regard. Various classification tasks, from the categorization of breast lesions to the detection of malignant thyroid nodules have thus already been improved by complementing a network's training stage with domain knowledge (Feng et al., 2020; Avola et al., 2021). The underlying study is to our knowledge the first to investigate this knowledge-driven GAN approach in MS imaging.

The value of machine learning (ML) tools generally hinges on their ability to remain accurate when deployed to data that is of different structure from the training data, making multicentric validation a mandatory prerequisite. Also, building trust in artificial intelligence (AI) is oftentimes hindered because the decision-making process is concealed to the user who can only accept or discard a binary output (Asan et al., 2020). Hence, providing visibility into how an ML system makes predictions has become a major concern, especially in the medical domain (Quinn et al., 2022). This can be achieved either by providing insights into the "black-box" problem of DL systems that are inherently uninterpretable by the human operator or by designing networks that are inherently interpretable but generally less potent (i.e., linear regression, decision-trees). Neural networks are a hallmark of the "black-box" problem as decisions are made through nonlinear associations between input and output, thus remaining opaque to the human reader. Improved interpretability can be achieved by decreasing the complexity of such networks (i.e., reducing the amount of neural connections), at the potential cost of performance loss, or through uncertainty measurements of the decision-making process (Le et al., 2020). By providing uncertainty maps that quantify the decision-making confidence of a GAN, the acceptance of synthetic MRI by the medical community might be improved while also offering clearer insights into potential causes for a system's malfunctioning. Uncertainty maps can be estimated by analysis of the variance across iterations during image synthesis, which has of late become an area of increasing interest (Gal and Ghahramani, 2015; Watson et al., 2019). Visualization of model confidence in GAN-mediated synthesis of MRI has been done before in tasks such as artificial motion-artifact inclusion or age prediction in fetal MRI (Shaw et al., 2020; Shi et al., 2020). In contrast to these works, we aim to quantify model confidence in translating areas of pathology that only constitute a small fraction of the generated data volume.

This study presents a refined GAN framework with an architecture that includes a task-specific training objective for MS lesion translation. We hypothesize that this GAN-based approach can provide synthetic, high-contrast DIR images from routinely acquired input FLAIR and T1w data, thereby removing the need for time-intensive acquisition of DIR. A special focus of this study is to evaluate this task-specific network for external validity in a multicenter dataset with scanners from different vendors and different acquisition details. To further provide an insight into the decision-making process of the GAN and guide the reviewing clinician toward potential artifacts, we

Abbreviations: MRI, magnetic resonance imaging; MS, multiple sclerosis; DIR, double inversion recovery; FLAIR, fluid-attenuated inversion recovery; GAN, generative adversarial network; DL, deep learning; ML, machine learning; AI, artificial intelligence; synthDIR, synthetic double inversion recovery; trueDIR, physically acquired double inversion recovery; SSIM, structural similarity index measure; LST, lesion segmentation tool; JC, juxtacortical; PV, periventricular; IT, infratentorial; SC, subcortical; LBR, lesion-to-background ratios; LFL, lesion-focused loss; NAWM, normal appearing white matter; PSNR, peak signal-to-noise ratio; ICC, intraclass correlation coefficient.

calculated uncertainty maps that reflect the variance in image-to-image translation.

MATERIALS AND METHODS

Patients

The study design was approved by the local IRBs and informed consent was obtained from all patients at their respective centers prior to scan acquisition.

Training Data

Data for model training were retrospectively retrieved from 50 patients with diagnosed MS and included T1w (2:28 min), FLAIR (3:55 min), and DIR (6:31 min). All scans originated from the same scanner (Ingenia 3.0T, Philips Healthcare, Best, Netherlands). Sequence parameters were identical in all patients for T1w (TR of 9.0 ms, TE of 4.0 ms, flip angle of 8°, acquired in the sagittal plane with an isotropic voxel size of 1 mm^3), FLAIR (TR of 4,800 ms, TE of 331 ms, TI of 1,650 ms, flip angle of 90°, acquired in the sagittal plane with an isotropic voxel size of 1 mm^3), and DIR (TR of 5,500 ms, TE of 355.9 ms, TI of 2,550 ms and 2,990 ms, flip angle of 90°, acquired in the sagittal plane with an isotropic voxel size of 1.1 mm^3).

Testing Data

Sixty MRI scans from 50 consecutive patients (20:20:10 for centers 1:2:3, respectively) with diagnosed MS were included. For centers 1 and 2, 1 scan/patient was sampled, while baseline and follow-up exams for 10 patients from center 3 were considered. MRI data included T1w, FLAIR, and DIR and were acquired on both, 3.0T and 1.5T scanners. In detail, testing data from center 1 was acquired on the same hardware and using the same protocol as the training data (Ingenia 3.0T, Philips Healthcare, Best, Netherlands), testing data from center 2 originated from a different 3.0T scanner from the same manufacturer (Achieva 3.0T, Philips Healthcare, Best, Netherlands), and testing data from center 3 was acquired on 1.5T and 3.0T scanners from a different manufacturer (Skyra 3.0T, Avanto_fit 1.5T, and Aera 1.5T, Siemens Healthineers, Erlangen, Germany).

Sequence parameters for T1w, FLAIR, and DIR sequences were chosen according to the site-specific parameters optimized for routine clinical imaging and not modified during the retrieval period (**Supplementary Table 1**). Dichotomization of data from centers 1–3 was made to acknowledge the fact that data structure from (1) corresponded to the training data (prospectively referred to as "internal data"), while the data structure from (2) and (3) was unknown to the network (prospectively referred to as "external data"). **Table 1** illustrates how the data was categorized for evaluation.

Double Inversion Recovery Image Synthesis

Network Architecture

Our GAN extends the existing "pix2pix" method (Isola et al., 2017) and is trained to synthesize a target image *y* (resembling the true target image *Y*) given a set of input images *X* and a lesion segmentation mask *S*. In this setting, two networks compete with each other: The generator *G* is based on a U-Net architecture and synthesizes the target DIR images (synthDIR) from two input images (T1w and FLAIR), while the discriminator *D* tries to determine if a given DIR image is synthetic (synthDIR) or physically acquired (trueDIR). The network architecture and training process of the GAN are given in **Figure 1**. Importantly, the input of T1 and FLAIR images are fed to U-Net to generate DIR images while the lesion mask is only used to compute additional lesion-specific loss during the training stage (see below). Thus the lesion segmentation mask *S* is not required during inference.

TABLE 1 | Data from center 1 was acquired on the same hardware as training data and thus considered to be of known structure (= internal data).

Data class (number of image sets)	Classes for study evaluation
Training data (*n* = 50)	
Test data from (1) (*n* = 20)	Internal data (Known data structure)
Test data from (2) (*n* = 20)	External data (Unknown data structure)
Test data from (3) (*n* = 20)	External data (Unknown data structure)

In analogy, data from centers 2 and 3 were acquired on different hardware and considered to be of unknown structure (= external data).

Loss Functions

The discriminator gives the judgment about how realistic the local structures are (called "Patch GAN"), and is patch-based and driven by a least-square error (L2) loss function (Mao et al., 2019). The generator is trained on a composite loss function based on (a) the reconstruction error between the synthesized image and the target image using SSIM and (b) the output of the discriminator when judging if a given image is either ground truth or synthetic. In addition to an SSIM, a peculiarity of our model is that an additional loss focusing on the successful translation of MS lesions was developed. In order to focus the model on MS lesions (which only make up a minority of voxels in an image), an additional L1 loss term is calculated between the true and synthetic DIR images after multiplying both images with the lesion segmentation mask *S*, thus only considering the translation of MS lesions for this part of the loss. The image reconstruction loss for the generator *G*, the loss function for the discriminator *D*, and the total loss function were formulated as follows, respectively:

$$\mathcal{L}_{recons} = 1 - SSIM\left(Y,\ G\left(X\right)\right) + \lambda_1 * \ ||(Y - G(X)) \odot S_1|| \tag{1}$$

$$\mathcal{L}_{\ D} = \mathbb{E}_X \left\{||1 - D(X)||_2\right\} \tag{2}$$

$$\mathcal{L}_{total} = \lambda_2 * \mathcal{L}_{recons} + \mathcal{L}_D \tag{3}$$

Here, λ_1 and λ_2 are hyper-parameters and set to 1 and 10, respectively, which balances the two loss components.

Optimization

The input and output images were co-registered, skull-stripped, linearly transformed into the MNI152 space, and resampled

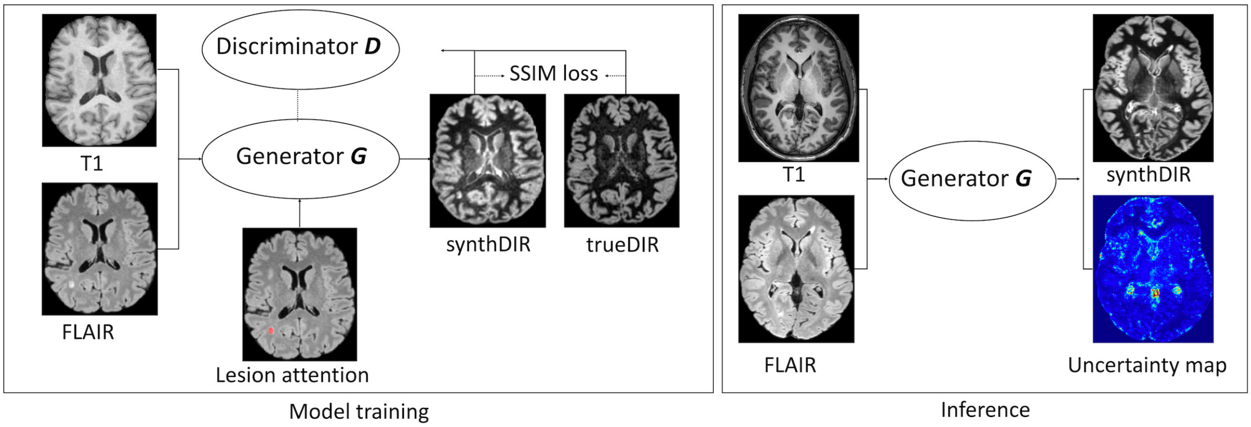


FIGURE 1 | Architecture, training process, and inference of the image synthesis task. The image Generator G uses the combination of FLAIR and T1w as input to generate synthDIR. The additional supervision from the lesion maps in the training stage drives an enhanced translation of MS-specific lesions (lesion attention). The feedback on the similarity between synthDIR and trueDIR is given by the Discriminator D and a structure similarity loss function and it updates the network weights until the loss function to discern both image pairs is minimal. During the inference stage, the trained generator G can generate the synthDIR and an uncertainty map showing the confidence of the output relating to each voxel. Uncertainty maps are calculated from the voxel-wise variances in signal intensities, as explained in the section "Materials and Methods".

to 1 mm isotropic resolution. As excellent correlation between automated and manual segmentation performance has been shown before, lesion segmentation maps were created using the Lesion Segmentation Tool (LST) (Schmidt et al., 2012). By including domain knowledge (in the form of lesion segmentation on FLAIR images) into the image translation during training, we enforced the model to pay attention to the lesion area by minimizing the difference between ground-truth images and synthetic images. In practice, such segmentation maps can be also provided by manual segmentation or other automated lesion segmentation tools (Schmidt et al., 2012; Li et al., 2018). Exemplary cases of all investigated sequences are shown in **Figure 2**. Training was carried out with a batch size of 1 for a total of 150 epochs, using the Adam optimizer with a learning rate of 0.001. During training, random intensity (gamma correction and gaussian blurring) and spatial (shifting and flipping) augmentations were performed. The best-performing model was selected using an internal validation set consisting of 10% of the training images.

The generated model is publicly available at https://figshare.com/articles/software/synthDIR/16607831.

Expert Readings

A dataset of 180 scans, comprising 60 sets each for FLAIR, synthDIR, and trueDIR, was investigated for lesion counts by two neuroradiologists (R1 with 5 years of experience in neuroradiological imaging, R2 with 3 years of experience in neuroradiological imaging) in a random order. Readers were blinded to scanner types and sequence labels. The number of juxtacortical (JC), periventricular (PV), infratentorial (IT), and subcortical (SC) lesions, in accordance with the 2017 McDonald criteria, were counted (Thompson et al., 2018). JC, PV, and IT lesions were considered to be MS-specific (Thompson et al., 2018). Albeit known to constitute different pathophysiological entities, we did not differentiate between cortical and juxtacortical lesions as this approach best reflects current guidelines (Bo et al., 2003; Thompson et al., 2018).

Quantitative Lesion Analysis and Uncertainty Maps

To quantitatively assess lesion translation, we calculated lesion-to-background ratios (LBR). Therefore, lesions on FLAIR and T1w images were segmented using LST, and tissue segmentation of T1w images was performed using ANTs Atropos (Avants et al., 2011). For comparison of LBR, GAN iterations with and without the above-stated lesion-specific loss function were computed.

From the segmentation maps, the lesion-to-background ratio was calculated as:

$$LBR = \frac{MeanSignal_{lesion}}{MeanSignal_{NAWM}} \tag{4}$$

Here, NAWM refers to "normal appearing white matter," i.e., non-lesioned white matter. From lesion segmentation maps and corresponding annotations in the NAWM, the mean signal intensity was extracted from DIR, FLAIR, and synthDIR images.

To estimate the GAN's uncertainty in generating synthDIR images, we performed variational inference during the test time by using dropout sampling. We added a dropout layer (dropout rate of 0.3) to the second-last layer of the U-Net and calculated 100 synthDIR images per input (Gal and Ghahramani, 2015). From these 100 iterations, we calculated the variance of voxel-wise intensities, resulting in the uncertainty map for visual inspection.

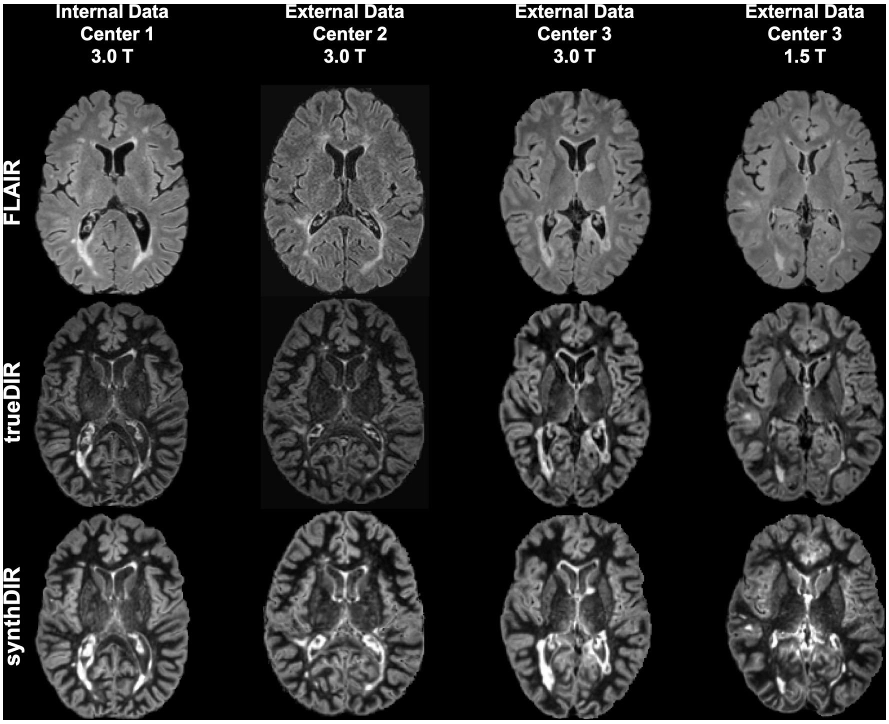


FIGURE 2 | Exemplary images of FLAIR, trueDIR, and synthDIR for all centers and scanners.

Statistical Analysis

Lesion counts were compared with a Wilcoxon signed-rank test to account for non-Gaussian distribution and paired data. LBR was compared with a paired *t*-test. Similarity of synthDIR and trueDIR was furthermore quantitatively assessed by the SSIM (Wang et al., 2004). For pixelwise comparisons, peak signal-to-noise ratio (PSNR) was calculated. Interrater agreement was assessed with the intraclass correlation coefficient (ICC; use of single measurements for absolute agreement in a two-way random model) and the related 95% confidence interval (95% CI). Statistical computations were performed with SPSS software (SPSS Statistics for Windows, version 25.0; IBM, Armonk, NY, United States). A *p*-value < 0.05 was considered statistically significant.

RESULTS

Interrater Agreement

Consistency between both readers was excellent with ICCs for all specific (JC + PV + IT) lesions ranging from 0.91 (95% CI: 0.85; 0.94) in FLAIR to 0.90 (95% CI: 0.84; 0.94) in synthDIR and 0.89 (95% CI: 0.83; 0.94) in trueDIR.

Lesion Counts

The study endpoint to improve depiction of MS specific lesions in synthDIR compared to FLAIR was met by both readers [26.7 ± 2.6 vs. 22.5 ± 2.2 ($p < 0.0001$) in R1 and 22.8 ± 2.2 vs. 19.9 ± 2.0 ($p = 0.0005$) in R2]. TrueDIR outperformed FLAIR in counts of MS-specific lesions [28.6 ± 2.9 vs. 22.5 ± 2.2 ($p < 0.0001$) in R1 and 23.3 ± 2.4 vs. 19.9 ± 2.0 ($p < 0.0001$) in R2]. While trueDIR remained superior to synthDIR in the depiction of MS-specific lesions in R1 [28.6 ± 2.9 vs. 26.7 ± 2.6 ($p = 0.0021$)], both image sets were of comparable diagnostic value in R2 [23.3 ± 2.4 vs. 22.8 ± 2.2 ($p = 0.98$)]. **Table 2** provides details on total and region-specific lesion counts for the study cohort.

Analysis of lesion counts as a function of scanner types revealed comparable effects independent of the structure of input data (internal or external). Hence, significant improvements in lesion counts were noted in synthDIR vs. FLAIR for both readers in external data [27.1 ± 3.4 vs. 22.6 ± 2.8 ($p < 0.0001$) in R1; 25.1 ± 2.9 vs. 21.5 ± 2.6 ($p = 0.0007$) in R2] and for R1 in internal data [26.6 ± 4.3 vs. 22.2 ± 3.6 ($p = 0.0029$) in R1; 18.1 ± 2.6 vs. 16.6 ± 2.6 ($p = 0.27$) in R2]. In external data, a slight improvement in lesion conspicuity was noted in trueDIR vs. synthDIR for R1 [28.9 ± 3.7 vs. 27.1 ± 3.4 ($p = 0.011$)] but not for R2 [25.6 ± 3.3 vs. 25.1 ± 2.9 ($p = 0.90$)]. **Table 3** provides lesion counts as a function of data source.

To increase the clinical reliability of synthDIR images, voxel-wise uncertainty maps from 100 forward runs using test-time dropout for Bayesian approximation were evaluated. For the majority of lesions, a high model confidence was observed, i.e., lesions were not highlighted in the uncertainty maps. On the other hand, artificial hyperintensities in synthetic images were readily identified by the high model uncertainty on

TABLE 2 | Lesion counts for all locations and both readers.

	All specific	P	PV lesions	P	JC lesions	P	IT lesions	P	SC lesions	P
Reader 1										
FLAIR vs. synthDIR	22.5 ± 2.2 vs. 26.7 ± 2.6	< 0.0001	12.0 ± 1.2 vs. 13.9 ± 1.4	< 0.0001	8.7 ± 1.2 vs. 10.8 ± 1.5	< 0.0001	1.9 ± 0.4 vs. 2.2 ± 0.4	0.043	10.6 ± 1.3 vs. 10.4 ± 1.2	0.82
FLAIR vs. trueDIR	22.5 ± 2.2 vs. 28.6 ± 2.9	< 0.0001	12.0 ± 1.2 vs. 13.9 ± 1.4	< 0.0001	8.7 ± 1.2 vs. 12.3 ± 1.7	< 0.0001	1.9 ± 0.4 vs. 2.4 ± 0.4	0.0002	10.6 ± 1.3 vs. 10.9 ± 1.4	0.36
SynthDIR vs. trueDIR	26.7 ± 2.6 vs. 28.6 ± 2.9	0.0021	13.9 ± 1.4 vs. 13.9 ± 1.4	0.91	10.8 ± 1.5 vs. 12.3 ± 1.7	< 0.0001	2.2 ± 0.4 vs. 2.4 ± 0.4	0.33	10.4 ± 1.2 vs. 10.9 ± 1.4	0.66
Reader 2										
FLAIR vs. synthDIR	19.9 ± 2.0 vs. 22.8 ± 2.2	0.0005	10.5 ± 1.0 vs. 12.4 ± 1.1	0.0004	7.8 ± 1.2 vs. 8.5 ± 1.3	0.18	1.5 ± 0.3 vs. 1.9 ± 0.3	0.024	13.5 ± 1.9 vs. 10.5 ± 1.5	< 0.0001
FLAIR vs. trueDIR	19.9 ± 2.0 vs. 23.3 ± 2.4	< 0.0001	10.5 ± 1.0 vs. 12.2 ± 1.2	0.0014	7.8 ± 1.2 vs. 9.7 ± 1.5	0.0028	1.5 ± 0.3 vs. 1.5 ± 0.3	0.99	13.5 ± 1.9 vs. 10.5 ± 1.6	< 0.0001
SynthDIR vs. trueDIR	22.8 ± 2.2 vs. 23.3 ± 2.4	0.98	12.4 ± 1.1 vs. 12.2 ± 1.2	0.26	8.5 ± 1.3 vs. 9.7 ± 1.5	0.068	1.9 ± 0.3 vs. 1.5 ± 0.3	0.03	10.5 ± 1.5 vs. 10.5 ± 1.6	0.70

PV, periventricular; JC, juxtacortical; IT, infratentorial; SC, subcortical; FLAIR, fluid-attenuated inversion recovery; trueDIR, real double inversion recovery; synthDIR, synthetic double inversion recovery.

TABLE 3 | Counts of MS-specific lesions for FLAIR, trueDIR, and synthDIR as a function of data source.

	All	P	Internal data	P	External data	P
Reader 1						
FLAIR vs. synthDIR	22.5 ± 2.2 vs. 26.7 ± 2.6	< 0.0001	22.2 ± 3.6 vs. 26.6 ± 4.3	0.0029	22.6 ± 2.8 vs. 27.1 ± 3.4	< 0.0001
FLAIR vs. trueDIR	22.5 ± 2.2 vs. 28.6 ± 2.9	< 0.0001	22.2 ± 3.6 vs. 27.9 ± 4.6	0.0001	22.6 ± 2.8 vs. 28.9 ± 3.7	< 0.0001
SynthDIR vs. trueDIR	26.7 ± 2.6 vs. 28.6 ± 2.9	0.0021	26.6 ± 4.3 vs. 27.9 ± 4.6	0.086	27.1 ± 3.4 vs. 28.9 ± 3.7	0.011
Reader 2						
FLAIR vs. synthDIR	19.9 ± 2.0 vs. 22.8 ± 2.2	0.0005	16.6 ± 2.6 vs. 18.1 ± 2.6	0.27	21.5 ± 2.6 vs. 25.1 ± 2.9	0.0007
FLAIR vs. trueDIR	19.9 ± 2.0 vs. 23.3 ± 2.4	< 0.0001	16.6 ± 2.6 vs. 18.6 ± 2.7	0.027	21.5 ± 2.6 vs. 25.6 ± 3.3	0.0001
SynthDIR vs. trueDIR	22.8 ± 2.2 vs. 23.3 ± 2.4	0.98	18.1 ± 2.6 vs. 18.6 ± 2.7	0.87	25.1 ± 2.9 vs. 25.6 ± 3.3	0.90

FLAIR, fluid-attenuated inversion recovery; trueDIR, real double inversion recovery; synthDIR, synthetic double inversion recovery.

these maps. **Figure 3** provides examples on how uncertainty maps allow to discern true-positive lesions from false-positive hyperintensities in synthDIR.

Quantitative Image Analysis

Similarity between trueDIR and synthDIR was highest in internal data, as shown by an SSIM of 0.967 ± 0.012, closely followed by external data (3) and (2) with still excellent SSIM-values of 0.950 ± 0.012 and 0.941 ± 0.010, respectively. For synthDIR, PSNR was highest in internal data at 29.2 ± 1.6 dB and decreased to 25.6 ± 1.1 dB in external data (3). **Table 4** provides detailed values for quantitative image metrics.

Effects of Lesion-Focused Loss Function

To assess the benefit of the lesion-specific loss function during image synthesis, LBR were compared between FLAIR, trueDIR, synthDIR, as well as synthDIR generated by a network iteration without the lesion-specific loss. Both versions of synthDIR, irrespective if additional loss was included or not, exceeded input FLAIR in LBR (data given in **Table 4**).

Of note, LBR was significantly lower in synthDIR generated by the version without lesion-focused loss compared to the version of synthDIR benefiting from lesion-focused loss (2.69 ± 0.66 vs. 2.80 ± 0.67, $p < 0.001$). While synthDIR achieved a comparable LBR to trueDIR (2.80 ± 0.67 vs. 2.86 ± 0.65, $p = 0.41$), this effect faded if synthDIR was generated without lesion-focused loss (2.69 ± 0.66 vs. 2.86 ± 0.65, $p = 0.032$) (as shown in **Figure 4**).

DISCUSSION

Medical imaging has benefited greatly from DL advances that gave birth to a panoply of systems aimed at tasks ranging from disease detection to image synthesis and artifact reduction (Emami et al., 2018; Rajpurkar et al., 2018; Liang et al., 2019). We here validated a GAN that has been fine-tuned to the translation of MS-specific white matter lesions while aiming to remain generalizable to external data. We further explored the concept of uncertainty maps to illustrate how trustworthy the network is in image-to-image translation. Such maps can

TABLE 4 | Image-wise (SSIM) and voxel-wise (PSNR) comparative metrics for synthDIR and trueDIR.

	SSIM (trueDIR – synthDIR)	PSNR (dB) (trueDIR – synthDIR)	LBR FLAIR	LBR trueDIR	LBR synthDIR	LBR synthDIR w/o LFL
All	0.954 ± 0.016	27.2 ± 2.2	1.52 ± 0.49	2.86 ± 0.65	2.80 ± 0.67	2.69 ± 0.66
Internal data	0.967 ± 0.012	29.2 ± 1.64	1.45 ± 0.06	2.80 ± 0.33	2.86 ± 0.34	2.68 ± 0.30
External data (2)	0.941 ± 0.010	25.8 ± 1.12	1.65 ± 0.12	3.01 ± 0.41	3.35 ± 0.50	3.31 ± 0.45
External data (3)	0.950 ± 0.012	25.6 ± 1.08	1.46 ± 0.86	2.78 ± 1.00	2.19 ± 0.56	2.07 ± 0.50

LBR are given for FLAIR, trueDIR, synthDIR, as well as for synthDIR generated by a GAN iteration without the lesion-focused loss function (synthDIR w/o LFL). Results are given for internal data, as well as external data (2) and (3). SSIM, structural similarity index measure; PSNR, peak signal-to-noise ratio; LBR, lesion-to-background ratio; LFL, lesion-focused loss; trueDIR, real double inversion recovery; synthDIR, synthetic double inversion recovery.

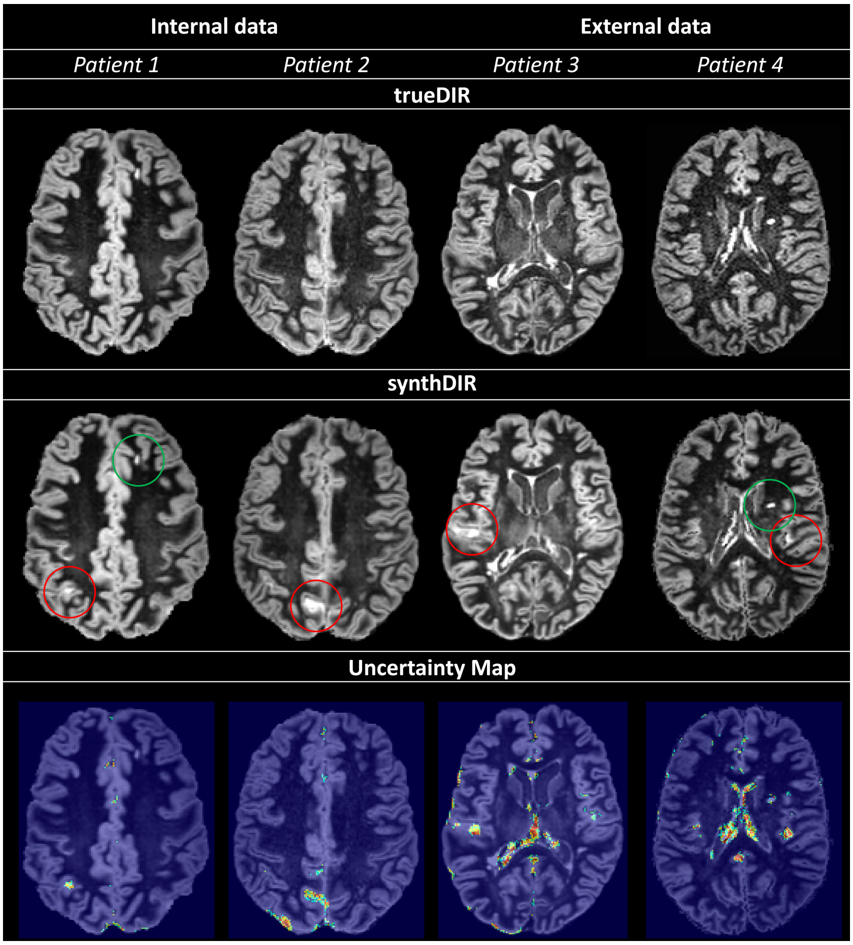


FIGURE 3 | Uncertainty maps provide relevant information regarding the validity of voxel-to-voxel translation; increases in uncertainty are scaled from blue to green. Circled in red (Patients 1–4) are hyperintensities in synthDIR without correlation in trueDIR and easily recognized as areas of high variance in the corresponding uncertainty maps, allowing for their identification as artifacts from the synthesis task. On the other hand, true-positive lesions are readily identified as regions with either no (patient 1 – green circle in synthDIR) or low (patient 4 – green circle in synthDIR) values of uncertainty. Hence, interpretation of synthDIR and decision-making on the veracity of lesions is facilitated through uncertainty maps.

provide important support to decide on the veracity of findings in synthetic images and help the radiologist to detect artifacts resulting from the synthesis task.

Comparison of the network's performance in internal and external data showed that significantly more MS-specific lesions could be found in synthDIR compared to the FLAIR sequence that was used as input, irrespective of the data origin. Approximately 20% more MS-specific lesions were thus depictable in synthDIR, a magnitude of difference that is of obvious clinical interest, especially in patients with low lesion counts. While other surrogates of MS activity have been explored, depiction of new inflammatory plaques is still considered the hallmark of disease monitoring in MS (Filippi et al., 2001; Chard et al., 2003; Wattjes et al., 2015). Also, lesion load has

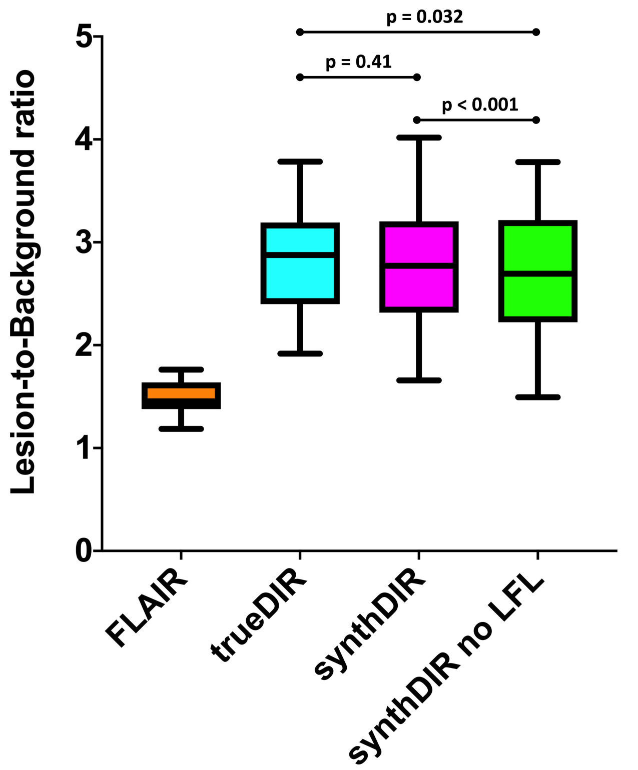


FIGURE 4 | Lesion-to-background ratios for FLAIR, trueDIR, and synthDIR. Additionally, LBR was calculated for synthDIR generated by a GAN-iteration without the lesion-focused loss (synthDIR no LFL). Of note, LBR was significantly higher in synthDIR compared to synthDIR without LFL, confirming the hypothesis that domain knowledge can be improved through LFL. While there was no significant difference in LBR between synthDIR and trueDIR (p = 0.41), LBR of synthDIR without LFL remained inferior to the LBR of trueDIR (p = 0.032). LBR, lesion-to-background ratio; LFL, lesion-focused loss.

been shown to directly correlate with future disability and, if properly detected and reliably quantified, might therefore prompt escalation of disease-modifying therapy (Calabrese et al., 2010; Popescu et al., 2011).

Domain knowledge, i.e., the ability to learn about pathology-specific image findings, is promising to further augment the clinical utility of DL tools (Yuan et al., 2019). The improved lesion translation that our GAN achieved by including a lesion-focused loss function hints at the potential of domain knowledge to further customize synthetic imaging. To highlight this, we showed that LBR in synthDIR was non-inferior to LBR in trueDIR only if the GAN was complemented by a lesion-focused loss.

The ability of synthDIR to outperform FLAIR, a sequence still considered gold-standard in MS imaging, has been shown for a multi-modal input (T1w, T2w, and FLAIR) in a monocentric setting (Finck et al., 2020; Bouman et al., 2021). In doing so, relevant reductions in scan times are feasible as the physical acquisition of 3D and isotropic DIR may take up to 7 min (Eichinger et al., 2019). While other methods, such as sparse sampling, have previously achieved scan time reductions for DIR, a GAN-based approach might be advantageous as it works on existing data and thus does not need to be prospectively deployed (Eichinger et al., 2019). This offers the potential advantage to augment the diagnostic value of existing studies and, hence, to render longitudinal follow-up exams more conclusive.

Albeit accurate in their output, neural networks generally fail to provide insight into the decision-making process, the so-called "black-box problem." Rendering this process more transparent is crucial for the acceptance of said networks and can, in theory, be achieved by providing methods to interpret the "black-box," or by designing models that are inherently more transparent in their functioning (Rudin, 2019; Arun et al., 2020). In GANs specifically, one potential bias in trying to match the (lesion) distribution in the target domain (trueDIR) is that features (lesions) with no correlation in source data might be erroneously mapped, a phenomenon commonly referred to as "hallucination." To verify lesion veracity we therefore introduced the concept of uncertainty maps that highlight the voxel-wise aleatoric variance taking place during image translation. Hence, the ability to compare hyperintensities in synthDIR to their respective uncertainty mappings can reduce the risk of false-positive findings, i.e., misinterpretation of constructed lesions in the synthetic image data. **Figure 3** illustrates how MS lesions can thus be separated from artifacts according to their voxel-wise intensity variance. As erroneous mappings remain an intrinsic limitation of GANs, their future deployment might benefit greatly from the calculation of uncertainty maps that are displayed in parallel to synthetic images.

A limitation of this approach is that having to reference synthDIR, along with the uncertainty maps adds complexity to the longitudinal interpretation of clinical MRI. Furthermore, comparison of synthDIR and trueDIR via autosegmentation techniques might have provided more objective lesion counts in this study. However, as our GAN was designed to provide synthetic data for clinical use, we opted for manual lesion counts as this best reflects the clinical reality. Future iterations of synthDIR might furthermore mitigate the wide disparities in lesion counts that we noticed especially in SC lesions. Also, prospective investigations should explore the feasibility to generate a GAN targeted to create synthDIR while using even fewer, potentially only one input modality. At last, we tested for generalizability by including three centers with differing hardware. Future investigations would benefit from the inclusion of more centers and readers, as our results demonstrate equivalence of synthDIR to trueDIR for only one of the two neuroradiologists.

CONCLUSION

Our findings confirm the use-case and external validity of a DL tool targeted at improving MRI in patients with MS. Our study demonstrates both, the utility of lesion-focused learning to improve domain adaption, as well as the potential benefit of uncertainty maps to help gain trust in GANs and make informed medical decisions. Presumably, wider deployment of these tools could prove beneficial in MS where treatment decisions are heavily relying on MRI findings.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of National Data Protection Law. Requests to access the datasets should be directed to TF, tom.finck@tum.de.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethikkommission Klinikum rechts der Isar. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BW, HL, BM, CZ, JK, TF, LG, and MM conceived and designed the project. BB and EB conceived the study and contributed external datasets. HL, BM, and BW designed the GAN. NS and SS performed the experiments. TF and BW prepared the writing of the first draft and performed the statistical analyses. TF prepared the figures and tables. All authors reviewed the first draft of the manuscript, contributed to the article, and approved the final manuscript draft.

FUNDING

BM, MM, and BW were supported by the DFG, SPP Radiomics. HL was supported by the Forschungskredit (Grant No. FK-21-125) from the University of Zurich.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2022.889808/full#supplementary-material

Supplementary Figure 1 | Network architecture: **(A)** The U-Net generator used to produce a synthetic DIR image from FLAIR and T1 input images. **(B)** The patch-based discriminator which receives as input both the source image (T1 and FLAIR) and either a real or synthetic DIR. The discriminator is patch-based. ks, kernel size.

REFERENCES

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., et al. (2020). Assessing the (Un)Trustworthiness of saliency maps for localizing abnormalities in medical imaging. *arXiv* [Preprint]. arXiv:2008.02766 doi: 10.1148/ryai.2021200267

Asan, O., Bayrak, A. E., and Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: focus on clinicians. *J. Med. Internet Res.* 22:e15154. doi: 10.2196/15154

Avants, B. B., Tustison, N. J., Wu, J., Cook, P. A., and Gee, J. C. (2011). An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics* 9, 381–400. doi: 10.1007/s12021-011-9109-y

Avola, D., Cinque, L., Fagioli, A., Filetti, S., Grani, G., and Rodolà, E. (2021). Multimodal feature fusion and knowledge-driven learning via experts consult for thyroid nodule classification. *IEEE Trans. Circuits Syst. Video Technol.* 1. doi: 10.1109/TCSVT.2021.3074414

Bo, L., Vedeler, C. A., Nyland, H. I., Trapp, B. D., and Mork, S. J. (2003). Subpial demyelination in the cerebral cortex of multiple sclerosis patients. *J. Neuropathol. Exp. Neurol.* 62, 723–732. doi: 10.1093/jnen/62.7.723

Bouman, P. M., Strijbis, V. I., Jonkman, L. E., Hulst, H. E., Geurts, J. J., and Steenwijk, M. D. (2021). Artificial double inversion recovery images for (juxta)cortical lesion visualization in multiple sclerosis. *Mult. Scler.* 28, 541–549. doi: 10.1177/13524585211029860

Calabrese, M., Filippi, M., and Gallo, P. (2010). Cortical lesions in multiple sclerosis. *Nat. Rev. Neurol.* 6, 438–444. doi: 10.1038/nrneurol.2010.93

Chard, D. T., Brex, P. A., Ciccarelli, O., Griffin, C. M., Parker, G. J., Dalton, C., et al. (2003). The longitudinal relation between brain lesion load and atrophy in multiple sclerosis: a 14 year follow up study. *J. Neurol. Neurosurg. Psychiatry* 74, 1551–1554. doi: 10.1136/jnnp.74.11.1551

Eichinger, P., Hock, A., Schon, S., Preibisch, C., Kirschke, J. S., Muhlau, M., et al. (2019). Acceleration of double inversion recovery sequences in multiple sclerosis with compressed sensing. *Invest. Radiol.* 54, 319–324. doi: 10.1097/RLI.0000000000000550

Emami, H., Dong, M., Nejad-Davarani, S. P., and Glide-Hurst, C. K. (2018). Generating synthetic CTs from magnetic resonance images using generative adversarial networks. *Med. Phys.* 45, 3627–3636. doi: 10.1002/mp.13047

Feng, H., Cao, J., Wang, H., Xie, Y., Yang, D., Feng, J., et al. (2020). A knowledge-driven feature learning and integration method for breast cancer diagnosis on multi-sequence MRI. *Magn. Reson. Imaging* 69, 40–48. doi: 10.1016/j.mri.2020.03.001

Filippi, M., Cercignani, M., Inglese, M., Horsfield, M. A., and Comi, G. (2001). Diffusion tensor magnetic resonance imaging in multiple sclerosis. *Neurology* 56, 304–311. doi: 10.1212/wnl.56.3.304

Finck, T., Li, H., Grundl, L., Eichinger, P., Bussas, M., Muhlau, M., et al. (2020). Deep-learning generated synthetic double inversion recovery images improve multiple sclerosis lesion detection. *Invest. Radiol.* 55, 318–323. doi: 10.1097/RLI.0000000000000640

Gal, Y., and Ghahramani, Z. (2015). "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning*, New York, NY. doi: 10.3390/s20216011

GBD 2016 Multiple Sclerosis Collaborators (2019). Global, regional, and national burden of multiple sclerosis 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* 18, 269–285. doi: 10.1016/S1474-4422(18)30443-5

Geurts, J. J., Pouwels, P. J., Uitdehaag, B. M., Polman, C. H., Barkhof, F., and Castelijns, J. A. (2005). Intracortical lesions in multiple sclerosis: improved detection with 3D double inversion-recovery MR imaging. *Radiology* 236, 254–260. doi: 10.1148/radiol.2361040450

Isola, P., Zhu, J., Zhou, T., and Efros, A. (2017). "Image-to-image translation with conditional adversarial networks," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI. doi: 10.1109/CVPR.2017.632

Le, V., Quinn, T. P., Tran, T., and Venkatesh, S. (2020). Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome. *BMC Genomics* 21(Suppl. 4):256. doi: 10.1186/s12864-020-6652-7

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W. S., et al. (2018). Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. *Neuroimage* 183, 650–665. doi: 10.1016/j.neuroimage.2018.07.005

Liang, K., Zhang, L., Yang, H., Yang, Y., Chen, Z., and Xing, Y. (2019). Metal artifact reduction for practical dental computed tomography by improving interpolation-based reconstruction with deep learning. *Med. Phys.* 46, e823–e834. doi: 10.1002/mp.13644

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P. (2019). On the effectiveness of least squares generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2947–2960. doi: 10.1109/TPAMI.2018.2872043

Popescu, B. F., Bunyan, R. F., Parisi, J. E., Ransohoff, R. M., and Lucchinetti, C. F. (2011). A case of multiple sclerosis presenting with inflammatory cortical demyelination. *Neurology* 76, 1705–1710. doi: 10.1212/WNL.0b013e31821a44f1

Quinn, T. P., Jacobs, S., Senadeera, M., Le, V., and Coghlan, S. (2022). The three ghosts of medical AI: can the black-box present deliver? *Artif. Intell. Med.* 124:102158. doi: 10.1016/j.artmed.2021.102158

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., et al. (2018). Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.* 15:e1002686. doi: 10.1371/journal.pmed.1002686

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1, 206–215. doi: 10.1038/s42256-019-0048-x

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Forschler, A., Berthele, A., et al. (2012). An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. *Neuroimage* 59, 3774–3783. doi: 10.1016/j.neuroimage.2011.11.032

Shaw, R., Sudre, C., Varsavsky, T., Ourselin, S., and Cardoso, M. J. (2020). A k-space model of movement artefacts: application to segmentation augmentation and artefact removal. *IEEE Trans. Med. Imaging* 39, 2881–2892. doi: 10.1109/TMI.2020.2972547

Shi, W., Yan, G., Li, Y., Li, H., Liu, T., Sun, C., et al. (2020). Fetal brain age estimation and anomaly detection using attention-based deep ensembles with uncertainty. *Neuroimage* 223:117316. doi: 10.1016/j.neuroimage.2020.117316

Thompson, A. J., Banwell, B. L., Barkhof, F., Carroll, W. M., Coetzee, T., Comi, G., et al. (2018). Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* 17, 162–173. doi: 10.1016/S1474-4422(17)30470-2

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 600–612. doi: 10.1109/tip.2003.819861

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., et al. (2019). Clinical applications of machine learning algorithms: beyond the black box. *BMJ* 364:l886. doi: 10.1136/bmj.l886

Wattjes, M. P., Lutterbey, G. G., Gieseke, J., Traber, F., Klotz, L., Schmidt, S., et al. (2007). Double inversion recovery brain imaging at 3T: diagnostic value in the detection of multiple sclerosis lesions. *AJNR Am. J. Neuroradiol.* 28, 54–59.

Wattjes, M. P., Rovira, A., Miller, D., Yousry, T. A., Sormani, M. P., de Stefano, M. P., et al. (2015). Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis–establishing disease prognosis and monitoring patients. *Nat. Rev. Neurol.* 11, 597–606. doi: 10.1038/nrneurol.2015.157

Yuan, W., Wei, J., Wang, J., Ma, Q., and Tasdizen, T. (2019). Unified attentional generative adversarial network for brain tumor segmentation from multimodal unpaired images. *arXiv* [Preprint]. arXiv:1907.03548

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

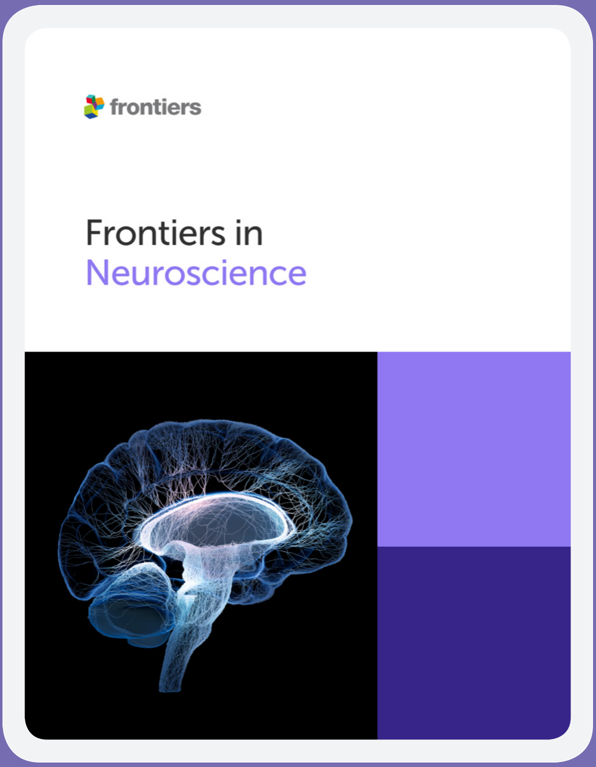
Frontiers in Neuroscience

Provides a holistic understanding of brain function from genes to behavior

Part of the most cited neuroscience journal series which explores the brain - from the new eras of causation and anatomical neurosciences to neuroeconomics and neuroenergetics.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Neuroscience

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact